NBER WORKING PAPER SERIES

NON-CHOICE EVALUATIONS PREDICT BEHAVIORAL RESPONSES TO CHANGES IN ECONOMIC CONDITIONS

B. Douglas Bernheim
Daniel Bjorkegren
Jeffrey Naecker
Antonio Rangel

Working Paper 19269
http://www.nber.org/papers/w19269

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2013

Previously circulated as "Do Hypothetical Choices and Non-Choice Ratings Reveal Preferences?" We would like to thank seminar participants at Stanford University's Behavioral and Experimental Economics Workshop, the 2011 ECORE Summer School (UCL, Louvain-la-Neuve), the 2012 ASSA Winter Meetings (Chicago), the 2012 CESifo Conference on Behavioral Economics (Munich), Harvard University, UCSD, and California State University, East Bay for helpful comments. Detailed suggestions from Richard Carson and Laura Taylor were especially helpful. We are also grateful to Irina Weissbrot for assistance with data collection. The first author also acknowledges financial support from the National Science Foundation through grant SES-1156263. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions
B. Douglas Bernheim, Daniel Bjorkegren, Jeffrey Naecker, and Antonio Rangel
NBER Working Paper No. 19269
August 2013, Revised March 2015
JEL No. C91,D12,H31,Q51

ABSTRACT

A central task in microeconomics is to predict choices in as-yet-unobserved situations (e.g., after some policy intervention). Standard approaches can prove problematic when sufficiently similar changes have not been observed or do not have observable exogenous causes. We explore an alternative approach that generates predictions based on relationships across decision problems between actual choice frequencies and non-choice subjective evaluations of the available options. In a laboratory experiment, we find that this method yields accurate estimates of price sensitivities for a collection of products under conditions that render standard methods either inapplicable or highly inaccurate.

B. Douglas Bernheim
Department of Economics
Stanford University
Stanford, CA 94305-6072
and NBER
bernheim@stanford.edu

Daniel Bjorkegren
Department of Economics
Box B
Brown University
Providence, RI 02912
USA
dan@bjorkegren.com

Jeffrey Naecker
Department of Economics
Stanford University
Stanford, CA 94305-6072
jnaecker@stanford.edu

Antonio Rangel
Department of Economics
California Institute of Technology
Pasadena, CA 91125
rangel@hss.caltech.edu

A data appendix is available at:
http://www.nber.org/data-appendix/w19269

1. Introduction

A central task in microeconomics is to predict households' choices in situations that have not yet been observed (e.g., after some proposed policy intervention). The dominant tradition is to draw inferences from actual choices within some related domain. Unfortunately, that approach often proves problematic due to various practical limitations of choice data.

If economic conditions vary (exogenously) in ways that closely resemble the intervention of interest within the set of decision problems for which choices are observed, and if it is possible to code those conditions quantitatively, one can infer the effects of the intervention using standard reduced-form methods. As an example, consider the canonical problem of estimating a demand curve based on price-quantity observations (assuming price variation is exogenous or an instrumental variable is available). In effect, one can predict demand at as-yet-unobserved prices through curve fitting and interpolation. When the aforementioned conditions are not met, one can sometimes use structural methods, in effect drawing inferences about likely effects from some form of variation in conditions that the structure implicitly deems "similar" to the intervention. As an example, consider the problem of predicting choices from non-linear budget sets when choices have only been observed for linear budget sets. A structural model of utility maximization tells us how the two are linked.

Of course, as the observed variation in economic conditions becomes further removed from the intervention of interest, the structural approach requires one to make ever-more-heroic assumptions concerning the relationships between the two.[2] In addition, the intervention of interest may have qualitative elements that are not easily incorporated into standard reduced-form or structural models. The latter concern arises regularly in behavioral economics, as studies have found that choices can depend critically on qualitative aspects of the decision frame both in the laboratory and in the field (see, e.g., Camerer et al., 2004, Bertrand et al., 2005, Saez, 2009). When we wish to predict choices in a novel decision frame, the reduced-form approach is inapplicable, while the structural approach requires a deeper structural understanding of the psychological processes that generate framing effects than, in most cases, we currently possess.

How can one proceed in contexts where standard reduced-form and structural methods prove problematic? We explore a novel strategy that employs data on subjective reactions to elements of contemplated opportunity sets when an individual is *not* engaged in making

[2] As an example, consider the problem of estimating the price elasticity of demand for health insurance among the uninsured, who are generally poor and not eligible for insurance through employers. After noting the difficulties associated with standard approaches (such as extrapolating from the choices of potentially non-comparable population groups), Krueger and Kuziemko (2013) turn to hypothetical choice data.

consequential choices. These *non-choice evaluations* include various types of subjective ratings (for example, measures of expected enjoyment, degree of temptation, and anticipated impact on specific objectives such as social image), as well as subjective "aggregators" such as stated preferences.[3] Our strategy presupposes that actual choices are observed in a number of distinct settings, but not the ones of interest. We predict choices for the latter in three steps. First, we elicit non-choice evaluations for all of the choice settings. Second, for the observed settings, we uncover the statistical relationships between these non-choice evaluations and actual choices. We study these relationships at the level of the choice problem, aggregating over decision makers. Using standard techniques, we select the relationships that prove most stable within the set of observed choice settings. Finally, we use those relationships to predict behavior for the settings of interest (which are out of sample). We discuss the conceptual reasons for thinking that such a strategy might be successful, and clarify its relation to existing methods, in subsequent sections. Because choice patterns (and hence preferences) are inferred from non-choice responses, we refer to this general class of procedures as *non-choice revealed preference* (NCRP).

We report the results of a laboratory experiment designed to gauge the potential usefulness of this approach. We offer subjects the opportunity to purchase a specified snack at a given price, such as $0.75, to be consumed during a waiting period. We then set the following task: supposing one only observes purchase frequencies at this price for all items (so that there is no observed price variation either for a single item or across items), can one accurately predict purchase frequencies for all items at a *different* price (such as $0.25)? Here, the price is intended to stand in for any economic condition (e.g., a policy) for which there is no usable historical variation (either because the policy has no close precedent, or because past policy variation is endogenous and there are no useful instruments).

Plainly, one cannot attack this task with standard reduced-form techniques, which require one to either interpolate or extrapolate from observed price variation (either within or across items). A conventional economic analysis might proceed by building a structural model (for example, in the spirit of Berry, Levinsohn, and Pakes, 1995), possibly one that infers the effect of price variation from the variation in serving size across items (which determines the price per gram), controlling for other differences. Alternatively, one might use stated preference techniques: simply ask people what they would choose at the alternative price, and take those

[3] See Shogren (2005, 2006), Carson and Hanemann (2005), and Carson (2012) for surveys of stated preference techniques. We discuss them at greater length in Section 7.

responses as indicating the actual purchase frequency. We show that those approaches work so poorly in this setting that they underperform a myopic benchmark (zero change in demand).

To implement our alternative approach, we use data on choices at the price that is assumed to have prevailed (e.g., $0.75) to estimate statistical relationships between real purchase frequencies and variables derived from non-choice evaluations (taking a choice problem as the unit of observation), and choose the best specifications based on within-sample selection criteria. We then use those relationships along with additional data on non-choice evaluations to predict real purchase frequencies at the alternative price (e.g., $0.25). The design of our experiment allows us to actually measure demand at both prices, and thus to gauge the accuracy of these predictions. We find that the specifications favored by within-estimation-sample model selection criteria predict purchase frequencies out of sample at the alternative price with a high degree of accuracy, both overall and across items. For example, starting from a price of $0.75, the average change in purchase frequency resulting from a change in price to $0.25 is +7.50 percentage points; the preferred bivariate specification predicts an average change of +7.66 percentage points, an error of just over 2%, without using any real purchase data at $0.25. More generally, the accuracy of our approach is well within the tolerances to which economists are accustomed, and its performance is roughly comparable to that of standard methods that require the analyst to observe within-item price variation for other items when projecting the demand for any given item at the alternative price. Accordingly, we conclude that NCRP methods have considerable potential.

The paper is organized as follows. Section 2 discusses the conceptual underpinnings for our approach. Section 3 provides details concerning the experiment. Section 4 sets forth the prediction task and the evaluation criteria. Section 5 identifies the benchmarks that we use to gauge predictive accuracy. Section 6 presents out main results. Section 7 clarifies the relationships between our methods and other existing approaches. Section 8 provides some concluding remarks, including a discussion of strategies for using this approach in practical applications. An online appendix contains extensive supplementary analyses.

2. Conceptual Framework

Suppose our object is to predict choices in some decision problem, D_0, that has not yet been observed. Ordinarily, an economist would seek data on choices made in a collection of related decision problems, $\{D_1, \ldots, D_N\} \equiv \mathbf{D}$, estimate models relating those choices to the objective characteristics of the problems, and then use those models to predict the choices people

would make in D_0. As noted in the previous section, this standard approach requires that the differences between the objective characteristics of D_0 and the elements of **D** are similar to the differences that exist within **D**. In many contexts, that condition is not satisfied.

As an alternative, we propose the following approach. First, gather new non-choice data pertaining to the decision problems D_0 and $D_1, \ldots, D_N$ (specifically, subjective evaluations pertaining to the available options, including various types of ratings and hypothetical decisions). Second, estimate models relating actual choices to variables derived from these non-choice evaluations using data for decision problems in **D**. Finally, use those models to predict choices in D_0. This alternative is similar to standard approaches, except that subjective characteristics of the decision problems (measured as non-choice evaluations) take the place of objective ones.

In some respects, the motivation for the proposed approach is simple. Although non-choice evaluations are known to suffer from significant biases, they are nevertheless strongly correlated with actual choices (see Sections 6 and 7). Presumably, they contain information pertaining to preferences, albeit in possibly biased, noisy, or otherwise distorted forms. Despite these distortions, such variables remain potentially useful as regressors in the types of models referenced above, provided the object is prediction rather than causal inference (see, e.g., White, 1980). For example, even though hypothetical choices often diverge systematically from actual choices, their high correlation with actual choices suggests that they may be good predictors, particularly when used in combination with other subjective evaluations that proxy for the sources of variation in hypothetical bias across decision problems.

The central advantage of the proposed approach is that the differences between the subjective characteristics of D_0 and the elements of **D** may be similar to the differences that exist within **D** even when the same statement does not hold for objective characteristics. For example, suppose that choices are motivated in part by perceptions of social approbation. The options available in D_0 may elicit roughly the same levels of approbation as some of the options available in $D_1, \ldots, D_N$ even if their objective characteristics are dissimilar. Consequently, our method may be applicable even when standard methods are not.

As with all standard methods for making out-of-sample predictions, the proposed approach requires stability of the predictive relationship over a domain that encompasses both the observed and not-yet-observed environments. The question of stability is best resolved through data analysis of the type we conduct in subsequent sections. However, as we explain next, there are good reasons to be optimistic about the prospects for finding adequately stable relationships.

Ultimately, choice depends on an individual's internal subjective representation of the available alternatives, rather than their objective characteristics (which are only relevant insofar as they influence the former). For any given decision problem, the individual perceives each alternative as addressing a collection of basic motivations to varying degrees, and these subjective perceptions map to choice (much as goods are valued for their "characteristics" in the seminal theory of Lancaster, 1966). Thus, the brain likely distills a (relatively) low-dimensional choice-relevant representation of each available option in terms of fundamental subjective attributes. If one could observe the subjective attributes associated with the available alternatives for a reasonably large collection of choice problems, one could recover the choice mapping.

As an illustration, consider decision problems involving large bundles of groceries. One can describe these bundles in a high-dimensional space of objective characteristics (e.g., by listing quantities of various products or ingredients), but for the purpose of predicting choices that representation is probably unhelpful. Alternatively, it might also be possible to describe these bundles in a low-dimensional subjective attribute space based on the degree to which each bundle is viewed as addressing basic motivations such as hunger, the desire for tasty food, health objectives, and image concerns. Critically, bundles with dissimilar objective characteristics may have similar fundamental subjective attributes. Accordingly, upon observing choices among a given set of bundles along with the pertinent subjective attributes, one may be able to predict choices among objectively dissimilar (but subjectively similar) bundles.

While the fundamental subjective attributes of choice options are not directly observable, it is relatively easy to elicit non-choice evaluations that pertain to those attributes. To illustrate, for our grocery example, one might conduct a survey in which respondents are asked to express the degree to which various bundles would address hunger, the desire for tasty food, health objectives, and image concerns, and then use those responses as regressors in prediction models.

Admittedly, responses to such questions are imperfect measures of the subjective attributes that actually determine choice. In some settings, people may be hesitant to say what they are actually thinking, and they may perceive or report motivations somewhat differently depending on whether they are actually making choices. But neither of those observations is necessarily fatal for the approach. Assuming the chosen alternative, x, is a function of a, the subjective attributes of available alternatives during choice (i.e., $x = C(a)$), and ignoring stochastic considerations for expositional simplicity, then as long as there is a stable relationship between a and non-choice evaluations n, i.e., $a = f(n)$, one can write choice as a function of those reactions: $x = \tilde{C}(n) \equiv C(f(n))$. (This notation obviously subsumes the case in which

$a = n$, but that assumption is not required.) Because our object is to make accurate choice predictions rather than to measure the causal effects of the subjective attributes accurately, nothing is lost by estimating $\tilde{C}$ rather than C.

The plausibility of assuming a stable relationship between a and n depends on context, but can be enhanced through a judicious selection of non-choice evaluations. For example, if one is concerned that complexity biases the assessment of subjective motivation, one can include subjective measures of complexity. Generally, to address potential divergences between actual and reported motivations, one can (a) include measures of the degree to which people say they believe others will understate or overstate an inclination in the context of a particular choice problem, (b) ask about the degree to which an alternative is likely to address motivations for peers, rather than for the respondent (thereby reducing incentives to misreport), and/or (c) employ non-choice reactions based on biometric responses, which are not subject to reporting bias.

More formally, it is natural to assume that non-choice reports of subjective attributes, a^r, depend on actual subjective attributes, and possibly on other context-specific considerations (z) that influence divergences between the two.[4] If we make the reasonable assumption that the tendencies to misreport z and a depend on the same set of motivational factors, we have $(a^r, z^r) = \big(g_a(a,z), g_z(a,z)\big) \equiv g(a,z)$ (where z^r stands for reported context-specific factors). As long as g is invertible, we can then write $a = f(n)$ where $n \equiv (a^r, z^r)$, and proceed by estimating $\tilde{C}(n)$, as above.

The approach outlined above presupposes stability of the functions C and g (and hence $\tilde{C}$) over a domain encompassing both **D** and D_0. A key part of that assumption – stability with **D** – is testable. (Stability over a domain that includes D_0 is not testable until choices for that decision problem are observed, but that is always true for out-of-sample predictions.) Indeed, when deciding whether to include particular non-choice reactions in n, one can take guidance not only from intuition and psychological research concerning motivational factors, but also from analysis of stability and cross-validated predictive accuracy within the estimation sample. For example, if observed conditions fall into several distinct regimes, one can test the stability of $\tilde{C}$ across regimes, and check the accuracy of predictions for each regime based on estimates of $\tilde{C}$ that employ data only from the other regimes.

[4] Actual subjective attributes may themselves influence these divergences. For example, if choosing a given alternative is viewed as likely to promote social esteem, then saying that it effectively addresses other motivations may also promote social esteem. Our approach allows for such possibilities.

A variety of factors likely determine the breadth of the domain over which $\tilde{C}$ is stable. One of these is the degree to which the implications for preferences of a given non-choice response is "portable" from one context to another. Depending on whether one is concerned with foods or paperback novels, knowing whether someone likes the taste of a given alternative has sharply different implications. Accordingly, this information is portable only with a limited domain, and its use would likely render a prediction model unstable outside that domain. In contrast, knowing whether social contacts would approve of a given alternative has similar implications irrespective of context. This information is highly portable, and a prediction model that only employs similarly portable variables is potentially stable over a broader domain.

A potential advantage of our approach is that the implications for preferences of appropriately chosen non-choice evaluations are much more portable than those of objective characteristics, and consequently predictive models are potentially stable over broader domains if they relate choices to the former and omit the latter. To illustrate, imagine estimating two competing models of the demand for potato chips, one employing physical characteristics such as crunchiness and salt content as regressors, the other employing non-choice evaluations such as ratings of tastiness. If these models are used to predict the demand for ice cream, the second will presumably perform far better than the first.

A "shortcut" strategy is to employ non-choice responses that aggregate over subjective attributes. One can think of a hypothetical choice as a non-choice reaction that putatively aggregates these dimensions comprehensively. However, it is well-established that hypothetical choices diverge systematically from actual choices.[5] As with other non-choice responses, we can plausibly assume that hypothetical choices (x^r) depend on motivational factors affecting real choices and reporting, $x^r = g_x(a, z)$, but there is good evidence that the mapping g_x differs from the real choice mapping C. Still, the use of hypothetical choice data potentially offers several advantages: it may economize on data collection (in the sense that one can use one aggregate in place of multiple components); it may encompass otherwise overlooked motivations; it is highly portable (in the sense that its meaning is identical in all contexts); and it is elicited through questions that respondents may find more familiar and easier to answer than inquiries about motivations.

The use of hypothetical choice data is plainly justified when posing a decision problem hypothetically only changes the scale of the measured behavioral response. Specifically, if we

[5] See, for example, Cummings et al. (1995), Johannesson et al. (1998), List and Gallet (2001), Little and Berrens (2004), Murphy et al. (2005), Blumenschein et al. (2007). When surveys are consequential, incentive problems also come into play; see Carson and Groves (2007) and Carson, Groves, and List (2011).

assume $g_x(a) = \tilde{x}(C(a))$ for some monotonic function $\tilde{x}$, then $x = \tilde{x}^{-1}(x^r)$, which we can estimate and use to predict choices. We can extend this reasoning to encompass the possibility that hypothetical choice is also distorted by a collection of context-specific factors, z. If we assume that $g_x(a,z) = \tilde{x}(C(a),z)$ (where $\tilde{x}$ is monotonic for each z), that $g_z(a,z) = \tilde{z}(C(a),z)$, and that $(x^r, z^r) = (\tilde{x}(x,z), \tilde{z}(x,z))$ is invertible, we then have $x = \tilde{C}(x^r, z^r)$ for some function $\tilde{C}$, which we can estimate and use to predict choices after the intervention.

In principle, one could apply this approach either at the individual level (to predict the decisions of a particular person), or at the aggregate level (to predict choice distributions). Because most economic applications are concerned with predicting the effects of interventions on the distribution of choices, we apply it here at the aggregate level. Doing so offers several other practical advantages. First, for an individual-level application, one would need to elicit both real choices and non-choice reactions regarding the same decision problems from the same individual, and it is likely that one type of response would contaminate the other (see Section 7, below). In contrast, for an aggregate-level application, one can gather data on real choices and non-choice responses from different samples of individuals without cross-contamination. Second, because this approach treats the decision problem as the unit of observation, precise estimation of $\tilde{C}$ requires data concerning both choices and non-choice reactions for a reasonably large collection of decision problems. One rarely has data pertaining to many distinct choice problems for a single individual, but distributions of choices for decisions made under differing conditions (and by different individuals) are readily available. Finally, individuals' idiosyncrasies likely average out in larger populations, facilitating more accurate prediction.

3. Experimental procedures and data

We tested this approach in a setting of intrinsic interest to economists: one in which the objective is to estimate demand curves for a collection of goods, but where no (usable) variation in price is observed. Because it is important for us to know the true price response for each good, we generated the data for this exercise through a laboratory experiment, which is free from the potentially spurious and confounding factors that often render the accurate measurement of such responses problematic in the field. In the interests of confronting our subjects with simple choices and evaluation tasks involving a reasonably large collection of familiar products, we settled on food items.

To gather data on the range of decisions and evaluations required for our analysis, we assigned each subject to one of multiple treatments, described below. At the outset of each

treatment session, subjects were told that the experiment would proceed in two stages. The first involved a computer-based choice or rating task lasting roughly 30 minutes. The second was a 30-minute waiting period. Subjects were not allowed to eat anything during the waiting period unless a snack was provided (according to the rules of the experiment). Sessions took place in mid-afternoon, when subjects are typically hungry.

In the first stage of each session, subjects either provided non-choice evaluations or made real decisions. The non-choice evaluations consisted of subjective ratings for some subjects and hypothetical choices for others. Real and hypothetical decisions pertained to snack food items offered at either $0.25 or $0.75. Subjective ratings pertained to the same collection of items, with price a factor in some but not all questions. For the reasons discussed in Section 2, we chose the subjective rating questions with the object of obtaining responses that contain information about the perceived effectiveness with which each alternative addresses underlying motivations, and/or about factors that may create divergences between real and hypothetical choices.

Each subject completed decision or non-choice evaluation task(s) for 189 snack food items (at both prices, where applicable), with the stimuli (food items or item-price pairs) presented in random order.[6] Subjects were divided into multiple task-specific treatment groups, with each subject participating in a single treatment to avoid cross-contamination of responses across tasks. Most treatment groups consisted of roughly 30 subjects. Altogether, 365 subjects participated (181 males, 184 females).[7] For a complete catalog of the treatment groups along with sample sizes and a screenshot for a representative question, see the Appendix A, section 1 and Figure A.1; the instructions used for each group appear in Appendix B. The following is a brief summary of key design features.

Some of the treatment groups provided subjective ratings. Depending on the group, subjects were asked to report their anticipated degree of happiness with each potential purchase, the anticipated degree of social approval or disapproval for each potential purchase, how much they liked each item, evaluations of regret, measures of temptation, expected enjoyment (ignoring considerations of diet or health), perceptions of health benefits, impact of consumption on social image, and the perceived inclination to overstate or understate the likelihood of a purchase.

[6] The items belong to the following eight broad categories: candy (48 items), cookies and pastries (40 items), chips and crackers (24 items), produce and nuts (18 items), cereal (14 items), drinks (11 items), soups and noodles (11 items), and other (25 items).

[7] We conducted the experiment at the Stanford Economic Research Laboratory (SERL). The protocol was reviewed and approved by Stanford University's IRB. Each subject was paid a participation fee between $20 and $30.We adjusted the fee upward when the response rate to our subject solicitation was low, and downward when it was high.

Several groups made hypothetical choices. The literature on stated preferences explores a variety of protocols for eliciting such choices, and attempts to determine which is most accurate. However, it is not clear that any single approach dominates the others. Indeed, it seems likely that different protocols elicit somewhat different (and potentially complementary) information. Accordingly, we employed multiple protocols, each with a separate treatment group. One protocol mimicked the real choice treatment (described below), except that no choice was implemented; we call this the "standard" protocol. A second protocol employed a "cheap talk" script (as in Cummings and Taylor, 1999) that encouraged subjects to take the hypothetical choices seriously,[8] a third elicited likelihoods rather than Yes/No responses (analogously to Champ et al., 1997), a fourth asked about the likely choices of same-gender peers (to eliminate image concerns and thereby potentially obtain more honest answers, analogously to Rothschild and Wolfers, 2011), and a fifth elicited willingness-to-pay (WTP) rather than Yes/No responses.

Finally, one group made real choices: one decision was selected at random and implemented during the 30-minute waiting period. A possible concern is that the low chance of implementing any given choice (one in 378 item-price pairs) renders it effectively hypothetical. Results presented in subsequent sections strongly refute this concern. Average purchase frequencies are significantly higher for hypothetical choices than for these real choices (consistent with the general finding in the literature concerning hypothetical bias); the cross-choice-task variance of the purchase frequency is considerably higher for hypothetical choices than for these real choices; and the average price sensitivity implied by the purchase frequencies is much larger for hypothetical choices than for these real choices. Plainly, despite the low implementation probabilities, subjects treated the real and hypothetical questions much differently.

It does not follow, however, that subjects viewed their "real" choices as entirely real, as opposed to partly real and partly hypothetical. To evaluate that possibility, we added a "mixed" treatment, in which subjects were told that five of their choices would be real (that is, one of the five would be chosen at random and implemented), and the rest would be hypothetical. The real choices were clearly identified and interspersed among the hypothetical ones. In that group, the implementation probability for each real choice was 1 in 5 rather than 1 in 378. We elicited 175 real choices through this "mixed" treatment, pertaining to 15 distinct items (at a price of $0.75). We then pooled that data with 450 choices involving the same 15 items from the "real choice" treatment, and estimated a probit regression relating the purchase decision to a set of 15 product dummies as well as a "mixed choice treatment" dummy. If the "real choice treatment" subjects

[8] We would like to thank Laura Taylor for generously reviewing and suggesting changes to the script, so that it would conform in both substance and spirit with the procedure developed in Cummings and Taylor (1999).

viewed their choices as real, the coefficient for the "mixed choice treatment" dummy should be zero; if they viewed those choices as partially hypothetical, then the "mixed choice treatment" coefficient should be negative given the documented direction of hypothetical bias. In fact, it was *positive* 0.0157 (probability scaled), with a standard deviation of 0.0364.[9] The difference is both statistically insignificant and of an economically small magnitude (1.57 percentage points). The coefficient indicates that the purchase frequencies were, if anything, slightly *higher* for real choices in the "mixed choice" treatment than in the "real choice" treatment, which is inconsistent with the hypothesis that participants in the "real choice" treatment were more inclined to view their choices as hypothetical than were participants in the "mixed choice" treatment.

We are not surprised by the finding that participants in the "real choice" treatment viewed their choices as real. After all, they had as much at stake as someone making a single purchase decision (because they knew one choice would definitely be implemented), and their task was no more tedious when taken seriously. Notably, similar conclusions were reached by Carson, Groves, and List (2011) based on theoretical principles and experimental evidence, and by Kang et al. (2011) based on fMRI data. Consistent with these findings, a survey paper by Brandts and Charness (2009) found no support for the hypothesis that differences between the strategy method and the direct response method increase with the number of contingent choices.[10]

4. Prediction task and evaluation criteria

We use the data gathered in our experiment to simulate the following empirical exercise. Suppose a large group of items (our 189 snack items) have all been sold only at a single price, P_1 (either \$0.25 or \$0.75), at which actual purchases have been observed. There is a proposal to change these prices to some new level, P_2 (\$0.25 if P_1 = \$0.75, and \$0.75 is P_1 = \$0.25). To help evaluate the proposal, an economist is asked to estimate the amount by which the demand for each of the items would increase or decrease. There is no opportunity to observe actual demand at any price other than P_1, but additional non-choice information is available.

As mentioned previously, this exercise is intended to stand in for any setting in which one wishes to estimate a behavioral response to a change in some economic condition, but either there is no observed variation in the condition, or the observed variation is not usable, perhaps because

[9] For the "mixed choice" treatment, purchase frequencies were significantly higher for hypothetical-choice items than for real-choice items (even though the choice frequencies for the two groups of items were very similar within the "real choice" treatment). Thus, the presence of real choices in the "mixed choice" treatment did not induce subjects to treat their hypothetical choices as real; they still suffered from hypothetical bias.

[10] It is important to acknowledge, however, that the pertinent studies involved far fewer contingent choices than in our "real choice" treatment.

it is endogenous and no valid instrument is available, or perhaps because it is qualitative and not easily reduced to a low-dimensional quantitative representation. For instance, the objective may be to gauge responses to a proposed policy change that has no close precedent. We have, of course, designed our experiment so that we observe actual choices in the setting of interest, and hence can evaluate predictive performance.

This prediction task fits within our conceptual framework as follows: D_n is a decision problem that offers item n at price P_1, while D_0 offers one of those items at price P_2.[11] Accordingly, our object is to predict the response of demand to a change in price based on the differences in demand across products sold at a single price.

A. Patterns of actual purchases

Before describing the criteria by which we evaluate the quality of predictions, it is important to verify first that our data on real choices manifests patterns that are worth predicting. Consequently, we begin by describing how the "real purchase frequency" (henceforth abbreviated RPF) varies across item-price pairs, of which there are 378 in total.

RPF varies from a low of 0 to a high of 60 percentage points, with a mean of 24.01%. Demand responds to price: the RPFs average 27.76% for a price of \$0.25 and 20.26% for a price of \$0.75 ($p \leq 0.001$).[12] As one would expect, the demand for these products is relatively price inelastic, but there is nevertheless a sizable average response (7.50 percentage points).

Conditional on price, the RPFs also vary considerably across items: the sample variance is 120.7 with a price of \$0.25 and 83.2 with a price of \$0.75. While these variances suggest that the attractiveness of the items varies considerably, it is important to bear in mind that, given the size of the "real choice" treatment group (30 subjects), some of that variation reflects sampling error. However, as we show in Appendix A, section 2, sampling error can only account for a fraction of the variation in RPFs across items; most of that variation is likely attributable to differences in underlying population frequencies.

There is also considerable variation across items in the responsiveness of the RPF to price changes; the variance of the percentage change in the RPF is 37.3. An increase in price from \$0.25 to \$0.75 reduces demand for 85.2% of our items, increases it for 2.6% of items, and has no effect for the remaining 12.2% of items. Much of the variation in the measured price response is presumably attributable to sampling error, which differencing may either amplify or reduce, depending on the magnitude of the correlation between choices by the same subject involving the

[11] Because we implement only one purchase decision, the demands for the products are independent by construction, and we can treat each such choice as a separate decision problem.
[12] Throughout, when comparing two means, we use paired t-tests.

same item but different prices. Without an estimate of that correlation, we cannot compute a useful bound on the fraction of the variance that is attributable to measurement error. However, in light of our ultimate success in generating predictions of price sensitivities that are reasonably well-calibrated (see Section 6), it is safe to conclude that some significant fraction of the variation in the measured responsiveness to price reflects population variation rather than sample variation.

B. Criteria for evaluating predictions

For each method of predicting demand at the new price (P_2) considered in subsequent sections, we evaluate the quality of predictions according to three criteria: overall bias, mean-squared prediction error (MSPE), and calibration.[13]

To be more specific, let $\hat{R}_{i2}$ denote the predicted RPF for item *i* at price P_2. Overall bias is absent, both for the predicted level of demand, $\hat{R}_{i2}$, and for the predicted change in demand, $\hat{R}_{i2} - R_{i1}$, when $\frac{1}{N}\sum_{i=1}^{N}\hat{R}_{i2} = \frac{1}{N}\sum_{i=1}^{N}R_{i2}$. The most common measure of overall bias, mean prediction error (MPE $= \frac{1}{N}\sum_{i=1}^{N}(\hat{R}_{i2} - R_{i2})$), has the limitation that magnitudes are not instantly interpretable. Instead, we calculate what we call the *normalized average effect* (NAE): $\frac{1}{N}\sum_{i=1}^{N}(\hat{R}_{i2} - R_{i1}) / \frac{1}{N}\sum_{i=1}^{N}(R_{i2} - R_{i1})$. A value of unity indicates no overall bias, a value such as 0.9 indicates that the predictions understate the actual effect of the price change by 10%, and a value such as 1.2 indicates a 20% overstatement.

We compute the MSPE for the level of predicted demand according to the standard formula MSPE $= \frac{1}{N}\sum_{i=1}^{N}(\hat{R}_{i2} - R_{i2})^2$. Notice that this is mathematically identical to the MSPE for the predicted change in demand, $\frac{1}{N}\sum_{i=1}^{N}((\hat{R}_{i2} - R_{i1}) - (R_{i2} - R_{i1}))^2$. Consequently, in what follows we will simply refer to the MSPE without specifying levels or changes.

Even a prediction that exhibits no bias on average may nevertheless be biased conditional on any given value of the prediction. As an extreme example, suppose the prediction is equal to the mean RPF across items, plus noise. In that case, the prediction would be unbiased on average, but biased conditional on it being any value other than its mean. We employ measures of calibration to address this issue. Specifically, if the predicted values of a variable are $\hat{X}_i$ and the actual values are X_i, we estimate a simple OLS regression of the following form:

$$X_i = \alpha + \beta\hat{X}_i + \varepsilon_i \qquad (1)$$

[13] As discussed below, we consider a model well-calibrated if, on average, realizations vary unit one-to-one with the model's forecasts. The term "calibration" is defined analogously in the statistical literature on probability models; see, e.g., Brier (1950) or Yates (1982). As noted in Section 7, "calibration" has an entirely different meaning in the literature on SP techniques.

If the prediction is perfectly calibrated (in the sense that $\hat{X}_i$ is an unbiased prediction of X_i conditional upon whatever value $\hat{X}_i$ takes for a given observation), then $\alpha = 0$, $\beta = 1$, and the conditional mean of ε is zero; thus, a simple OLS regression should yield these values. The parameter β is of particular interest because it governs the manner in which bias varies with the value of the prediction. In contrast, α pertains only to the average bias (which NAE also measures). Therefore, we report β as our measure of calibration for the predictions $\hat{X}_i$.

Our calibration parameter is *not* mathematically equivalent for predicted levels of demand, $\hat{X}_i = \hat{R}_{i2}$, and predicted changes in demand, $\hat{X}_i = \hat{R}_{i2} - R_{i1}$. Therefore, we report both. As we will see, the task of achieving good calibration is typically more challenging for predicted changes in demand than for predicted levels.

It is also important to provide a gauge of precision, so that one can assess whether the extent of any departure from an ideal standard, or any apparent improvement of one prediction method over another, is greater than one would expect to observe based on chance alone. Accordingly, we derive a joint bootstrap distribution for all prediction methods and evaluation metrics by drawing 1000 resamples at random from the set of items. The literature raises questions about the use of bootstrap procedures in contexts involving the types of variable selection methods we employ (see, e.g., Leeb and Pötscher, 2005, and Kyung et al., 2010), but there appear to be no workable alternatives as of this writing.[14] The reader should bear this qualification in mind when interpreting statements about the bootstrap distribution pertaining to the predictive performance of those methods.

5. Benchmarks

When evaluating the quality of a prediction, it is important to have in mind benchmarks that help one gauge "good" performance. We employ two classes of benchmarks. The first involves prediction methods that employ no more data on actual choices than the methods we wish to evaluate. Given the ground rules of our exercise, these methods are *feasible* alternatives to our approach; one could also use them in the applications we envision. The second class involves methods that require additional data on actual choices. These methods are *infeasible* under our ground rules; one could not actually deploy them in the same applications. We consider them because they establish demanding benchmarks.

A. Benchmarks that use limited choice data

[14] See West (2006) for a survey of forecast evaluation methods.

If, as our prediction task assumes, choice data are limited to RPFs for all of our items at a single price, P_1, the options for predicting RPFs at the alternative price, P_2, without using non-choice data are limited. The first line of Table 1 provides performance statistics for the simplest alternative, a myopic prediction (no price response, $\hat{R}_{i2} = R_{i1}$). We do not report a calibration statistic for the predicted change in demand because it is zero for all items. We view the myopic benchmark as a minimal standard: any approach that underperforms it is not worth considering.

We cannot use reduced-form methods to generate a benchmark using limited choice data because those data are assumed to exhibit no variation in price either within or across items. Structural methods require one to observe variation in some condition that is "similar" to price variation according to an assumed structural model. Here, the natural candidate is variation in serving size (quantity) across items (controlling for the items' characteristics), because a difference in quantity implies a difference in price per unit.

To construct a structural model, we assume that subject s derives value $V_i + \varepsilon_{is}$ from good i, where V_i is a component of item i's value common to all subjects, and ε_{is} is an iid random variable with standard error σ such that $\frac{\varepsilon_{is}}{\sigma} \equiv \eta_{is} \sim$ Logistic(0,1). Because we use the strategy method, the purchase decisions for all goods are independent, so subject s buys item i iff $V_i + \varepsilon_{is} \geq P_1$ (where P_1 is the single price charged for all items). We also assume that $V_i = (X_i\beta)q_i$, where X_i is a vector of characteristics, β is a vector of parameters, $X_i\beta$ is the value of the item per gram, and q_i is the number of grams.[15] The latter assumption imposes two restrictions on the common value component: first, it is zero when quantity is zero; second, it is linear in quantity. The first restriction is reasonable; the second is defensible given the small quantities involved. We note that this assumption is critical for the identification of price effects, since otherwise a doubling of price would not be equivalent to a halving of quantity. Details concerning the estimation of this model, as well as its use for forecasting, appear in Appendix A, section 3.

As shown in the second line of Table 1, even the most accurate structural model we examined performs terribly out of sample. The model implies average price responses nearly three times as large as the actual responses, so the overall bias is nearly twice the actual price response. It also performs worse than the myopic benchmark in terms of MSPE. Calibration for levels is acceptable when predicting from \$0.75 choices to \$0.25 choices, but not when predicting

[15] Alternatively, we could use another variable (such as calories) rather than grams as the scaling factor for utility. If we defined q_i to be the number of calories, $X_i\beta$ would be the value of the item per calorie. The characteristics X_i might then include grams per calorie. In practice, using grams rather than calories as the scaling factor yielded the best-performing predictive model.

from \$0.25 choices to \$0.75 choices, and calibration for changes is quite poor (particularly when predicting from \$0.75 choices to \$0.25 choices).

To be clear, we do not interpret this finding as a general indictment of structural methods. Rather, it shows that the prediction task we have set ourselves is a challenging one. That is why the success of our method, documented below, is notable.

Another feasible alternative under our ground rules is to employ stated preference techniques, treating hypothetical purchase frequencies (henceforth abbreviated HPF) as predictions rather than as predictors. We use two methods to predict the RPFs at the price P_2. The first is simply to set $\hat{R}_{i2} = H_{i2}$ (where H_{ij} denotes the HPF for item i at price j); we call this the "levels method." The second is to set $\hat{R}_{i2} = R_{i1} + (H_{i2} - H_{i1})$; we call this the "difference method." One would expect the difference method to outperform the levels method when the forecast errors for the latter are highly correlated within item across prices (e.g., if the degree of hypothetical bias is an item-specific fixed effect).

Not surprisingly, the data exhibit substantial hypothetical bias: the average standard-protocol HPF (30.88%) overstates the average RPF (24.01%) by nearly 7 percentage points (equivalently, by 28.6%), and we reject the absence of bias ($p < 0.001$). Moreover, the HPF exceeds the RPF for 73% of item-price pairs. Nevertheless, there is a strong correlation across items between the RPF and the HPF ($\rho = 0.697$), which suggests that the HPF may be a useful predictor of the RPF, even if it is not a good prediction.

The variance of the HPF is more than twice that of the RPF, a phenomenon we call *hypothetical noise*.[16] One might conjecture that this pattern emerges because hypothetical choices are more random than real choices, possibly as a result of subjects taking them less seriously. However, that explanation is incorrect. As we show in Appendix A, section 4, hypothetical noise is attributable in significant part to greater systematic variability of *population* HPFs than of *population* RPFs across choice problems. A possible explanation is that, when answering hypothetical questions, people naturally exaggerate the sensitivity of their choices to pertinent conditions; for example, as noted below, our data exhibit this pattern with respect to price variation. This result is encouraging: if we can identify the characteristics of choice problems that account for the sizable difference in variation between population HPFs and RPFs, we will be in a position to construct vastly improved forecasts of RPFs for unobserved choice problems.

[16] Similarly, Carson, Groves, and List (2011) found that the variance of valuations rises when choices become less consequential.

Together, hypothetical bias and hypothetical noise render the standard-protocol HPF a remarkably poor prediction of the RPF, regardless of whether one uses the differences method or the levels method; see the third and fourth lines of Table 1. For the difference method,[17] the predicted price response is nearly twice the actual response, so the overall biases of the difference method and the myopic benchmark are roughly equal. The MSPE is substantially larger than for the myopic benchmark, and calibration in levels is noticeably poorer. The calibration parameter for changes, for which the myopic benchmark provides no counterpart, is extremely low (0.248). The levels method generally underperforms the difference method, the exceptions being NAE and MSPE when predicting from \$0.25 to \$0.75.

In evaluating calibration results based on OLS regressions, it is important to bear in mind that we measure HPFs and RPFs for groups of modest sizes, rather than for the population. Even if the relationship between the RPF and an HPF reflects perfect calibration for the population, it will not do so in a finite sample, because the sample HPF measures its population counterpart with error.[18] In particular, the distribution of H_i conditional on H_i^P (the population HPF) is binomial with mean H_i^P. Whether one should worry about the implications of that observation depends on one's objective. If the objective is to assess calibration conditional on measuring the HPF and the RPF with groups of a particular size, the OLS regression provides the pertinent information. But if the objective is to assess the calibration one could achieve by using sufficiently large groups, the OLS estimates are contaminated by errors-in-variables (EIV) bias.

To gauge large-group calibration, we consider two alternative measures of calibration for changes using the differences approach, and of calibration for levels using the levels approach.[19] For the first, we reduce sampling error by doubling the size of the sample used to compute the HPFs. For the second, we re-estimate equation (1) using our original sample, but instrument for HPF using the HPF measured with the duplicate sample. Because the HPFs for the original and duplicate samples reflect the same population tendencies, they are necessarily correlated, and because they reflect independent random draws from the population, their sampling errors are

[17] For the difference method, the absolute value of the average bias is the same when predicting from \$0.75 choices or from \$0.25 choices; only the sign changes. Likewise, the MSPE and calibration for differences are exactly the same in either case. However, calibration for levels differs according to whether we are predicting \$0.75 or \$0.25 choices,

[18] Sampling error in the measurement of the RPF should not affect calibration for levels using the levels approach or calibration for changes using the differences approach because, in those cases, an RPF appears only on the left-hand side of the regression equation. However, such sampling error will affect calibration for changes using the levels approach and calibration for levels using the differences approach, because in those cases an RPF also appears on the right-hand side of the regression equation.

[19] Even though we can estimate the variance of the measurement error using the properties of the binomial distribution, we cannot compute the magnitude of the EIV bias by applying the standard formula, because (a) the variance of the measurement error, and hence the noise-to-signal ratio, varies according to the true value of the HPF, and (b) given our procedures, the measurement error is likely correlated across item-price pairs.

necessarily uncorrelated either with population HPFs or with each other. Consequently, the IV approach should yield a consistent estimate of the calibration parameter for the population.

With the difference method, our measure of calibration for changes rises from 0.248 to 0.312 when we double the sample, and to 0.519 when we instrument. With the levels method, predicting $0.25 choices, our measure of calibration for levels rises from 0.474 to 0.528 when we double the sample, and to 0.688 when we instrument; predicting $0.75 choices, our measure of calibration for levels rises from 0.466 to 0.516 when we double the sample, and to 0.671 when we instrument. Thus, increases in group size can improve calibration, but only to a limited degree.

As we mentioned in Section 3, studies in the literature on stated preferences gather hypothetical choice data using a variety of protocols, some of which are intended to "fix" the standard hypothetical choice question. While we find that some of the alternative protocols reduce the overall degree of hypothetical bias compared with the standard protocol, it appears that they generally do so in our experiment by introducing offsetting biases, rather than by addressing the underlying cause of the bias. Strikingly, the overall correlation between the RPF and the standard-protocol HPF is higher than for any alternative HPF, which casts doubt on the hypothesis that the alternative protocols improve the informational content of the hypothetical choice measures. With one exception, "3rd party" hypothetical choices by same-gender peers, none of the alternatives yields a clear improvement over the standard hypothetical choice protocol, and even that approach performs quite poorly compared to the benchmarks that use additional choice data, discussed in the next subsection. See the Appendix A, section 5, for details.

As shown in the table, in most cases the 0.5th-to-99.5th percentile intervals of the bootstrap distribution exclude the value of unity (the ideal) for NAE and the various calibration metrics we use to evaluate the methods discussed in this subsection.

B. Benchmarks that also use additional choice data

For our second set of benchmarks, we assume that purchase decisions are observed for the item of interest at the price P_1, and for other items at *both* prices, P_1 and P_2 . Thus, one can use the behavioral response to price for other items to predict the response for the item of interest. In practice, we randomly divide the items into five "folds" of (approximately) equal sizes, and forecast the RPFs at the price P_2 for items in each fold (the "hold-out sample") assuming that the

available choice data encompass price variation for all items in the other four folds (the "training sample," consisting of 80% of the items).[20]

We examine four benchmarks of this type. First, we simply compute the mean change in the RPF (i.e., $R_{i2} - R_{i1}$) for the items in the training sample, and predict R_{k2} for each item in the holdout sample by adding that average response to R_{k1}. Even with no adjustment for differences across products, this benchmark provides a reasonably demanding standard, as it presupposes that one can observe a wealth of data describing behavioral responses to price variation for closely related items, contrary to the ground rules governing our main prediction task.

For the remaining three benchmarks, we use the training samples to estimate models of the form

$$R_{i2} = \alpha + \beta R_{i1} + X_i\gamma + \varepsilon_i \,, \qquad (1)$$

and employ these models to predict R_{i2} for items in the hold-out sample. One version omits X_i; for the others, X_i includes variables that identify food categories and measure nutritional context. We use OLS because it has desirable forecasting properties (see, e.g., White, 1980). However, we also recognize that OLS is susceptible to the overfitting problem in contexts where the number of potential predictors is large relative to the number of observations. Because overfitting can compromise the accuracy of out-of-sample predictions, we also employ LASSO (the Least Absolute Shrinkage and Selection Operator, due to Tibshirani, 1996), a widely used technique from machine learning. As the name implies, LASSO is a *shrinkage estimator*, which means it compensates for overfitting by shrinking the overall size of the coefficient vector. Shrinkage can attenuate the sensitivity of predictions to changes in predictors, and hence reduce variance, thereby improving the accuracy of out-of-sample predictions according to measures such as mean-squared prediction error.[21]

Measures of predictive performance appear in the bottom portion of Table 1. All of these approaches yield substantial improvements over the myopic benchmark. Much of the gain is

[20] When measuring calibration for these benchmarks, we introduce fold fixed effects into the pertinent regressions, so that the performance metric is determined by correlations between predictions and realizations within folds. Correlations between predictions and realizations across folds are potentially spurious: mechanically, when the average of the outcome variable is higher in the holdout sample, it is lower in the estimation sample.

[21] Formally, LASSO optimizes a standard criterion for within-sample fit (here, it minimizes the sum of squared residuals) subject to a penalty that is proportional to the size of the normalized coefficient vector, measured in the L_1-norm (i.e., the sum of the absolute normalized coefficients). The form of the penalty ensures that LASSO assigns coefficients of zero to many potential regressors, thereby performing variable selection as well as shrinkage. Because of the variable selection feature, we allow LASSO to draw on a larger set of product characteristic variables than we employ for any particular structural model. A list of the included variables for the LASSO-selected specifications in Table 1 appears in section 7 of the Appendix.

achieved simply by assuming that the price response for each item would be the same as the average response for other items.[22] Allowing the prediction to be conditioned more flexibly on the value of the RPF at the price P_1 yields some improvement when predicting from \$0.25 choices, but not when predicting from \$0.75 choices. Predictive performance actually deteriorates when the model is augmented to include the items' characteristics. This finding reflects the general principle that parsimonious models often predict better than ones with large numbers of apparently relevant variables. However, the LASSO procedure, which pares down the list of predictors and shrinks the coefficient vector to combat overfitting, yields a meaningful improvement over the other approaches. Specifically, it achieves the lowest values of MSPE, and near-ideal values for NAE and all but one calibration parameter.

The benchmarks described in this subsection provide demanding standards for evaluating methods of forecasting price responses in settings where no price variation is observed for any item. Because they involve standard and widely used methods, any approach that achieves comparable results using markedly inferior data ought to merit serious consideration.

6. Results

Next we evaluate the accuracy of predictions based on statistical models relating RPFs to non-choice responses, including subjective ratings and HPFs (elicited with various protocols). As in Section 5C, one can construct these predictions using either the levels method or the difference method. For the levels method, we simply set $\hat{R}_{i2} = R_{i2}^F$ (where R_{i2}^F denotes the fitted value of $\hat{R}_{i2}$ based on the model); for the difference method, we set $\hat{R}_{i2} = R_{i2} + \left(R_{i2}^F - R_{i1}^F\right)$. For the sake of brevity, we report results based only on the difference method, which almost always outperformed the levels method in practice.

A key step in building good predictive models is model selection. In the current context, the set of possible models is enormous because we can potentially draw on an extremely large set of predictors. Specifically, we have assessed a variety of non-choice reactions, and in the case of hypothetical choices have used a number of protocols. For questions that can elicit more than two distinct responses, we have a separate variable measuring the frequency of each response (leaving one out because the frequencies sum to one). Non-linear terms (such as squares of average responses) and interactions may also prove predictively useful.

The criteria used for model selection must pertain to performance *within the training sample*; it would not be valid to evaluate our approach by selecting models that yield the best out-

[22] For this benchmark, we do not report calibration for changes, because the prediction does not vary within each fold.

of-sample predictions. As mentioned in Section 5.B, the LASSO procedure was devised (in part) to assist with model selection in settings where the objective is accurate out-of-sample prediction, and where the number of potential predictors is large relative to the number of observations. Consequently, our first step is to select and estimate a model using LASSO, allowing it to draw on the entire set of variables constructed from non-choice responses.[23]

The second row of Table 2 reports our metrics of out-of-sample predictive accuracy for the resulting LASSO models. The NAE and calibration parameters are all reasonably close to unity, and in five of six cases the 10^{th}-to-90^{th} percentile intervals of the bootstrap distribution do not exclude unity (the ideal value). Remarkably, even though our approach employs no data on actual choices at the new prices, it exhibits modest overall bias: it understates the average effect of a price change by 15% when predicting from \$0.75 to \$0.25, and overstates it by about 8% when predicting from \$0.25 to \$0.75. Improvements over the myopic benchmark with respect to NAE and MSPE are dramatic.[24]

In the first row of the table, we also reproduce results for the best-performing benchmark (which employs choice data encompassing price variation, and is therefore infeasible according to our ground rules). Overall, the LASSO specifications perform well compared with that benchmark: they achieve lower MSPE when predicting from \$0.75 to \$0.25 and overall (averaging across the two directions), as well as better overall calibration in changes. Calibration in levels is only slightly worse than the benchmark when predicting from \$0.75 to \$0.25 (0.948 versus 1.023 for the benchmark); though it is noticeably worse when predicting in the opposite direction,[25] the calibration coefficient remains reasonably high (0.759).

We can potentially achieve further improvements by fine-tuning our model selection criteria. One technique is to seek specifications that perform well in cross-validation exercises. For cross-validation, one simulates out-of-sample predictive performance by dividing the training sample into folds, and treating each fold (one at a time) as the hold-out sample. Instead of assigning observations randomly to multiple folds, we divide the observations into two folds according to whether the value of the HPF in the "duplicate" sample (discussed in Section 5.C) is above or below the median.[26] To understand why, recall that our out-of-sample predictions either

[23] Lists of the included variables for all of the LASSO-selected specifications in Table 2 appear in section 7 of the Appendix.
[24] In each case, the 0.5^{th}-to-99.5^{th} percentile interval of the bootstrap distribution for the difference in performance excludes zero.
[25] In this case, the 5^{th}-to-95^{th} percentile interval of the bootstrap distribution for the difference in performance excludes zero.
[26] Results based on random of assignment of observations into multiple folds are qualitative similar, though out-of-sample predictions are typically a bit less accurate.

employ data for a relatively attractive group of alternatives (snacks priced at $0.25) to forecast choices for a relatively unattractive group of alternatives (snacks priced at $0.75), or the other way around. Because the duplicate HPF captures aspects of an option's attractiveness (aside from price), dividing the training sample into folds according to the value of the HPF allows us to simulate the predictions of interest more closely than random assignment to multiple folds.

We fine-tune model selection by using a hill-climbing algorithm to identify the OLS model that optimizes specified measures of cross-validated predictive performance within the training sample. We initialize each search with a model that employs the same variables as the LASSO specification. We conduct one search to find the model that optimizes cross-validated performance according to each of our evaluation criteria -- overall bias, MSPE, and calibration.[27]

Metrics of out-of-sample predictive accuracy for the resulting models also appear in the top section of Table 2.[28] We begin with specifications that maximize the quality of cross-validated calibration, because that aspect of predictive performance appears to pose the greatest challenge. All performance metrics improve compared with the LASSO specification when predicting from $0.25 choices, but three of four deteriorate when predicting from $0.75 choices (the exception being NAE). Averaging across both directions, MSPE improves slightly. The specifications that minimize the cross-validated magnitude of overall bias produce a qualitatively similar pattern of improvement and deterioration, but underperform the specifications that maximize cross-validated calibration quality in 7 of 8 cases. The specifications that minimizes cross-validated MSPE generally underperform the others, but nevertheless yield almost no overall bias when predicting from $0.75 choices.

It is natural to wonder whether the accuracy of our approach hinges on employing data on large numbers of subjective response variables capturing *both* hypothetical choices *and* other types of subjective ratings. As alternatives, we examined the predictive power of simple models relating the RPF to the HPF elicited with a single protocol, as well as to two HPFs elicited with different protocols. Details appear in Appendix A, section 6; here we highlight some key findings.

We selected among the various univariate and bivariate models based on two standard (within-estimation-sample) criteria. One is the AIC (Akaike Information Criterion), a measure of

[27] For calibration, we focus on levels rather than differences. For both NAE and calibration, we search for the specification that minimizes the distance from unity.

[28] Lists of the included variables for all specifications in Table 2 that are optimized through cross-validation appear in section 7 of the Appendix.

goodness-of-fit that includes a penalty based on the number of parameters in the model,[29] which is commonly used for model selection when accurate out-of-sample prediction is the objective.[30] The second criterion is cross-validated MSPE. Among univariate models, these criteria generally favor a specification that relates the RPF to the standard-protocol HPF. Apparently, the alternative elicitation methods emphasized in the literature do not improve the informational content of hypothetical choices, at least in this setting. Among bivariate models, the same criteria often favors a specification that relates the RPF to HPFs based on the standard and "3rd party" protocols.

Metrics of out-of-sample predictive accuracy for the preferred univariate and bivariate models appear in the middle section of Table 2. Overall, these models perform surprisingly well. The average biases implied by the NAEs are quite small. The univariate model understate the average effect of a price change by 7.3% when predicting from $0.75 choices, and by 5.8% when predicting from $0.25 choices. The bivariate model is even more accurate on average: it overstates the average effect of a price change by 2.2% when predicting from $0.75 choices, and understates it by 2.5% when predicting from $0.25 choices. In terms of MSPE, both specifications outperform the myopic benchmark by a wide margin and achieve a substantial fraction of the improvement associated with the multivariate specifications.[31] Calibration for the simple models is respectable in levels, but considerably weaker in differences; the multivariate specifications achieve the greatest gains with respect to the latter metric.[32]

The preferred univariate model yields accurate predictions because the statistical relationship between the RPF and the standard HPF does not depend to any significant extent on price. Based on a Chow test, one cannot reject the hypothesis that the regression coefficients are the same for observations involving items sold at a price of $0.25, and for those involving items sold at a price of $0.75 ($p$ = 0.593). Figure 1 shows why. We have used orange dots for item-price pairs with prices of $0.25, and blue dots for pairs with prices of $0.75. Visually, lowering the price appears to shift the cloud to the northeast without disturbing the relationship between the variables. To drive this point home, we have plotted separate regression lines for the $0.25

[29] When comparing specifications with the same number of predictors, rankings of specifications by the AIC coincide with rankings by R^2, but that is not the case when comparing specifications with different numbers of a predictors.
[30] Results based on another well-known alternative, the BIC (Bayesian Information Criterion) are similar.
[31] For both the bivariate and univariate models, the 0.5th-to-95.5th percentile intervals of the bootstrap distributions for the improvements in NAE and MSPE over myopic predictions excludes zero.
[32] Focusing first on the improvement in this calibration parameter when moving from the univariate to the LASSO specifications, the 2.5th-to-97.5th percentile interval of the bootstrap distributions excludes zero when predicting from $0.75 choices, and the 10th-to-90th percentile interval excludes zero when predicting from $0.25 choices. Focusing next on the improvement in this calibration parameter when moving from the bivariate to the LASSO specifications, the 5th-to-95th percentile interval of the bootstrap distributions excludes zero when predicting from $0.75 choices.

choices and the $0.75 choices on the figure, along with error bands. For all practical purposes, they are indistinguishable.[33] Similarly, a Chow test reveals no significant differences between the $0.25 and $0.75 versions of the bivariate model ($p = 0.937$).

We also estimated LASSO specifications that draw on all of the hypothetical choice variables, but none of the other subjective ratings. Results also appear in the middle portion of Table 2. Curiously, the LASSO procedure selects models that perform considerably worse than the best univariate and bivariate models in terms of overall bias and MSPE, even though it could in principle replicate those models. This finding underscores the difficulty of identifying within-sample model selection criteria that assure good out-of-sample predictive performance. A specification that draws only on the ratings variables and none of the hypothetical choice variables performs even less well.

Next we ask whether one can achieve further improvements in predictive accuracy by adding the physical characteristics of the items to the list of potential predictors.[34] It is important to emphasize that any such improvements potentially come at a cost. Ideally, the research agenda set forth in this paper would eventually identify predictive statistical relationships that are stable over reasonably broad domains, so that one can extrapolate likely behavior from hypothetical choices and non-choice ratings without gathering sufficient data to estimate highly context-specific predictive models. For that purpose, it is important to use predictors for which implications concerning preferences do not vary over the intended domain. For the most part, we have focused on non-choice reactions for which these implications are largely independent of the domain – e.g., how much a subject likes an outcome, how happy they would be with it, the extent to which others would approve, which they would choose, and so forth. In contrast, the implications of physical characteristics can vary dramatically across domains. For example, greater sugar content may be a desirable characteristic for chocolate, but not for mustard. Consequently, by employing objective characteristics, we may improve predictive power within some narrow domain, but impair the model's applicability outside that domain.

Results for LASSO specifications that draws on all variables measuring hypothetical choices, non-choice ratings, and physical characteristics appear in the bottom section of Table 2. There are modest improvements compared with the original LASSO model. The NAEs are

[33] To determine whether our finding is driven by the use of linear functional forms, we reestimated the relationships nonparametrically using kernel regression. Though there is a bit of weaving back and forth, the two curves remain virtually on top of each other (see Figure A.2 in Appendix A).

[34] The characteristics are as follows: calories, calories from fat, fat (g), sodium (mg), carbohydrates (g), sugar (g), and protein (g), all per serving, as well as category dummies for drinks, candy, produce & nuts, cookies & pastries, chips & crackers, cereal, soup & noodles, and uncategorized.

indicative of slightly smaller overall bias: with this approach, we understate the average effect of a price change by 11.3% when predicting from \$0.75 choices, and overstate it by 3.0% when predicting from \$0.25 choices. The overall MSPE, averaged over the two directions, is slightly lower, and three of the four calibration parameters are closer to unity, but the differences are modest. Notably, with respect to the most challenging performance metric (calibration in differences), this approach uniformly outperforms the most demanding benchmark.

These findings admit at least two interpretations. One is that a context-specific prediction model can potentially perform somewhat better than a fully portable one. The other is that any sacrifice in predictive accuracy resulting from eschewing context-specific variables to enhance the model's portability across domains is relatively small.

We have seen that the accuracy with which one can predict the price response of an item is roughly the same when "good" choice data are available (i.e., we observe choices at different prices for closely related items), so that one can estimate specifications in the form of equation (1), and when no price variation is observed but non-choice response variables are available. We close this portion of our analysis by asking whether the addition of non-choice response variables improves predictive accuracy even when one has access to good choice data. The final line of Table 2 contains results for LASSO estimates of a specification in the form of equation (1), where the vector X_i is augmented to include not only product characteristics, but also a full set of variables measuring hypothetical choices and non-choice ratings. Relative to a specification that omits the latter variables (results for which appear in the first row of the table), performance noticeably improves with respect to both MSPE and calibration for changes, and there is no significant sacrifice in other dimensions.[35] Thus, the use of non-choice response variables significantly enhances predictive performance even when good choice data are available.

7. Related Literature

Our approach is related to stated preference (SP) techniques and the contingent valuation method (CVM), which make extensive use of hypothetical choice data (for reviews, see Shogren, 2005, 2006, Carson and Hanemann, 2005, and Carson, 2012). This literature seeks to predict choices for non-market goods when choice data pertaining to closely related decisions are *entirely* unavailable (e.g., in the environmental context, to value non-market goods such as pristine

[35] Notably, this is the only model for which the ideal value (unity) lies within the 10th-to-90th intervals of the bootstrap distributions of both NAEs and all four calibration parameters.

coastlines);[36] in contrast, we explore the use of non-choice data as an alternative or supplement to choice data even when the latter are available (but are not ideal).[37]

It is well-established that answers to standard hypothetical questions are systematically biased.[38] Two classes of solutions have been examined. One attempts to "fix" the hypothetical question.[39] Our approach is more closely related to a second class of solutions involving *ex post* statistical calibration.[40] These techniques exploit statistical relationships between real and hypothetical choices and, like our approach, treat the latter as a predictor rather than a prediction.

The *ex post* calibration techniques used in the SP/CVM literature differ from ours in several ways. The main distinguishing feature of our approach is that it treats the decision problem as the unit of observation and relates choice distributions to the problem's (subjective) characteristics. In contrast, *ex post* calibration techniques treat the individual as the unit of observation and relate hypothetical bias to his or her socioeconomic and demographic characteristics. While those techniques account for differences in hypothetical bias across individuals (for a given decision problem), they cannot account for differences across decision problems. Consequently, they are not useful for predicting choice distributions in decision problems that have not yet been observed.[41] On the contrary, List and Shogren (1998, 2002) emphasize that hypothetical bias is context-specific, so that individual-level calibration does not reliably transfer from one setting to another.[42] Yet psychological studies also suggest that

[36] In some cases, the object is to shed light on *dimensions* of preferences for which real choice data are unavailable by using real and hypothetical choice data in combination; see, e.g., Brownstone et al. (2000) and Small, Winston, and Yan (2005).

[37] Studies that use non-choice data as an alternative and/or supplement to choice data even when the latter are available (but are not ideal) are relatively rare. As an example, consider the problem of estimating the price elasticity of demand for health insurance among the uninsured, who are generally poor and not eligible for insurance through employers. One possibility is to extrapolate from the choices of potentially non-comparable population groups, which also requires one to grapple with the endogeneity of insurance prices, as in Gruber and Washington (2005). Alternatively, Krueger and Kuziemko (2011) recently attacked the same issue using hypothetical choice data, and reached strikingly different conclusions (i.e., a much larger elasticity).

[38] The bias typically favors overstatement of willingness-to-pay and alternatives that are viewed as more "virtuous." See, for example, Cummings et al. (1995), Johannesson et al. (1998), List and Gallet (2001), Little and Berrens (2004), Murphy et al. (2005), Blumenschein et al. (2007). When surveys are consequential, incentive problems also come into play; see Carson and Groves (2007) and Carson, Groves, and List (2011). Biases do not appear to be substantial in all settings, however; see, for example, Abdellaoui, Barrios, and Wakker (2007) for a within-subject comparison of choices over lotteries and stated (cardinal) preferences over monetary payments.

[39] Methods include the use of (1) certainty scales (as in Champ et al., 1997), (2) entreaties to behave as if the decisions were real (as in the "cheap-talk" protocol of Cummings and Taylor, 1999, or more recently the "solemn oath" protocol of Jacquemet et al., 2010), and (3) "dissonance-minimizing" protocols (as in Blamey et al., 1999, and Loomis et al., 1999, which allow respondents to express support for a public good while also indicating a low WTP).

[40] See Kurz (1974) Shogren (1993), Blackburn, Harrison, and Rutstrom (1994), National Oceanographic and Atmospheric Association (1994), Fox et al. (1998), List and Shogren (1998, 2002), and Mansfield (1998).

[41] Indeed, unlike our analysis, existing ex post calibration studies do not generally focus on out-of-sample predictive performance. Nor do they run the types of "horse races" between choice-based and non-choice-based prediction methods that reveal whether these methods have merit in settings where (imperfect) choice data are also available.

[42] Blackburn et al. (1994) provide somewhat mixed evidence on portability, but their analysis is limited to two goods.

hypothetical bias is systematically related to measurable factors that vary across decision problems (e.g., Ajzen et al., 2004, and Johansson-Stenman and Svedsäter, 2003). Our approach allows us to adjust for factors affecting the degree of hypothetical bias that vary across decision problems by including other appropriate non-choice responses, such as questions that elicit norms or image concerns.

An additional advantage of conducting our analysis at the level of the decision problem is that we can assess non-choice responses using different groups of subjects. In contrast, in *ex post* calibration studies, subjects make real choices after making hypothetical ones, which introduces the possibility of cross-contamination.[43] Our ability to obtain independent non-choice responses with distinct groups also allows us to employ, in a single specification, combinations of predictors that include multiple versions of hypothetical choices (e.g., standard, certainty scaled, and cheap-talk variants) along with other subjective ratings, and to determine whether those measures have independent and complementary predictive power. In contrast, the aforementioned studies calibrate hypothetical choices one version at a time.

A separate pertinent strand of research within the SP/CVM literature involves meta-analyses (Carson et al., 1996, List and Gallet, 2001, Little and Berrens, 1994, and Murphy et al., 2005). Unlike the *ex post* calibration literature, those studies attempt to find variables that account for the considerable variation in hypothetical bias across contexts and goods. However, they are primarily concerned with evaluating the effects of diverse experimental methods on hypothetical bias,[44] rather than with assessing out-of-sample predictive accuracy, as we do.

Stepping away from SP data, portions of the neuroeconomics literature seek to predict choices from neural and/or physiological responses. Smith, Bernheim, Camerer, and Rangel (2014) focus specifically on passive non-choice neural reactions, and provide proof-of-concept that those types of reactions predict choices.[45] Separately, in the literature on subjective well-being, two recent papers explore the relationships between forward-looking statements concerning happiness and/or satisfaction and *hypothetical* choices (Benjamin et al., 2010, 2012), which motivates our use of such variables to predict *real* choices.

Turning to other disciplines, the marketing literature has examined stated intentions as predictors of purchases (see, e.g., Juster, 1966, Morrison, 1979, Infosino, 1986, Jamieson and

[43] While Blackburn et al. (1994) do not reject the hypothesis of no contamination, their test is limited to a single setting and its power is unclear. Moreover, marketing studies have found, on the contrary, that stated intentions influence subsequent choices (see, e.g., Chandon et al., 2004, 2005). Similarly, voter surveys have been shown to affect turnout (see, e.g., Kraut and McConohay, 1973).

[44] One exception is that they point to a systematic difference in hypothetical bias for public and private goods.

[45] See also Tusche et al. (2010) and Levy et al. (2011).

Bass, 1989). Its relationship to our work is similar to that of the SP/CVM literature on *ex post* calibration techniques in that the object, once again, is to derive individual-specific predictions for a given good, with cross-good differences addressed through meta-analysis (e.g., Morwitz et al., 2007). Marketing scholars also routinely use SP data (derived from "choice experiments" involving hypothetical choices over multiple alternatives) to estimate preference parameters in the context of a single choice problem (see Louviere, 1993, Polak and Jones ,1993, Ben-Akiva et al., 1994, or Alpizar et al., 2003, for a useful review). Our analysis provides methods for potentially improving those data inputs. There are also parallels to our work in the political science literature, particularly concerning the prediction of voter turnout and election results, e.g., from surveys and polls (as in Jackman, 1999, and Katz and Katz, 2010). As in our approach, the object is to predict aggregate outcomes rather than individuals' choices, and a range of potential predictors (in addition to hypothetical choices or intentions) are sometimes considered. For example, Rothschild and Wolfers (2011) find that questions concerning likely electoral outcomes (i.e., how *others* will vote) are better predictors than stated intentions.[46] The problem is substantively different, however, in that surveys and polls ask voters about *real* decisions that many have made, plan to make, or are in the process of making, instead of measuring non-choice reactions to choice problems that respondents view as hypothetical.

8. Concluding remarks

We have reported the results of a laboratory experiment designed to evaluate the potential usefulness of methods involving non-choice revealed preference, and to compare their accuracy with conventional approaches. Hypothetical choice frequencies are poor predictions of real choice frequencies, but are nevertheless good predictors, particularly when used in combination and with other non-choice ratings. Consequently, using NCRP methods, it is possible to forecast the effect on demand of a change in price, even if no usable price variation is observed.

This paper is properly construed as only the start of a research agenda. Much work remains. As we have seen, issues involving model selection can be especially thorny, and merit further examination. Significantly, in other contexts, blind adherence to mechanical within-sample selection criteria may be inappropriate. It is possible, for example, that such criteria would discard a variable measuring the most important dimension of motivation differentiating the environments of interest from the observed environments, simply because other motivational

[46] Some studies also use prediction markets (e.g., Rothschild, 2009), which (in effect) elicit investors' incentivized forecasts of electoral outcomes.

factors vary more, and consequently play more important predictive roles, within the latter. In such cases, variable selection must be guided in part by a conceptual understanding of the ways in which the environments differ. One possible solution would be to perform model selection subject to a constraint that certain presumptively relevant variables must be included.

Other unexplored issues concern the breadth of the domain over which predictive relationships are usefully portable, and the related issue of how much context-specific accuracy must be sacrificed to achieve greater portability. We are also far from exhausting the range of subjective questions that might yield valuable predictors.

If the methods explored in this paper are to be of practical value, it will be necessary to resolve various pragmatic and conceptual issues concerning their use in actual applications. One potential strategy for applying these methods would operate as follows. Imagine that the object is to estimate the effect on choice of various policy options, and to predict the effects of a untried policy with novel elements. Using naturally occurring data, one could estimate a regression relating behavior to coarse features of the existing policies, but this would not permit one to extrapolate the effects of the untried policy. (Moreover, if the observed policies are qualitatively differentiated, it might be difficult to represent them in a regression format.) Instead, one could recruit a fixed subject pool and elicit subjective reactions concerning the extent to which the observed and untried policy regimes would impact motivations to behave in particular ways. These responses would then be aggregated across subjects and used as measures of the policies' "subjective attributes." Using these measures in combination with the naturally occurring data, one could then estimate a regression relating behavioral outcomes to the subjective attributes of the policy environments (in effect replacing measures of the policies' objective attributes with subjective ones). The predictive validity of such models could be assessed through cross-validation, possibly holding out sub-classes of policies one at a time. If validated, the model could then be used to predict behavior in the untried policy environment based on its measured subjective characteristics. We are currently exploring several applications along these lines.

In some real-world contexts, nominally hypothetical questions are either consequential or perceived as such, and consequences do not incentivize truthful revelation. For example, when asked about the frequency with which they would likely fly a new route, airline customers who expect to use that service have incentives to exaggerate. Though we have focused here on prediction from inconsequential responses, our methods are also potentially applicable to improperly incentivized responses (though the predictive relationships would likely be different).

At this stage in our research, we have not sought a structural understanding of the processes governing the relationships between real choices and non-choice responses. Through structural modeling, one could potentially obtain predictive models that are stable across domains of even greater breadth. Whether one takes a non-structural approach (as in this paper) or a structural one, a potential advantage of this strategy over conventional methods of predicting choices in as-yet unobserved situations is that it may ultimately require an understanding of only a single process (one determining how choices are related to the fundamental subjective attributes of the available alternatives), rather than a distinct model for every decision context.

References

Abdellaoui, Mohammed, Carolina Barrios, and Peter P. Wakker, "Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory," Journal of Econometrics 138, 207, 356-378.

Ajzen, Icek, Thomas C. Brown, and Franklin Carvajal, "Explaining the Discrepancy between Intentions and Actions: The Case of Hypothetical Bias in Contingent Valuation," *Pers Soc Psychol Bull* 30, 2004, 1108-1121.

Alpizar, Francisco, Fredrick Carlsson, and Peter Martinsson, "Using Choice Experiments for Non-Market Valuation," *Economic Issues* 8(1), 2003, 83-110.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao, "Combining Revelaed and Stated Preferences Data, " *Marketing Letters* 5(4), 1994, 335-350.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and A. Rees-Jones, "Do People Seek to Maximize Happiness? Evidence from New Surveys," NBER Working Paper 16489, 2010.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot, "Beyond happiness and satisfaction: developing a national well-being index based on stated preference," mimeo, 2012.

Berry, Seven, James Levinsohn, and Ariel Pakes, "Automobile prices in market equilibrium," *Econometrica* 63(4), 1995, 841-890.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman, "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," NBER Working Paper No. 11892, December 2005.

Blackburn, McKinley, Glenn W. Harrison, E. Elisabet Rutström, "Statistical Bias Functions and Informative Hypothetical Surveys," *American Journal of Agricultural Economics* 76(5), December 1994, 1084-1088.

Blamey, R. K., J. W. Bennett, and M. D. Morrison, "Yea-saying in contingent valuation surveys," *Land Economics* 75(1), 1999, 126–141.

Brandts, Jordi, and Gary Charness, "The Strategy versus the Direct-response Method: A Survey of Experimental Comparisons," mimeo, November 9, 2009.

Brier, G. W. , "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, 78 (1950), 1-3.

Brownstone, David, David S. Bunch, and Kenneth Train, "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles," *Transportation Research* 34, 2000, 315-338.

Camerer, Colin F., George Loewenstein, and Matthew Rabin (eds.), *Advances in Behavioral Economics*, Princeton, NJ: Princeton University Press, 2004.

Carson, Richard T., "Contingent Valuation: A Practical Alternative when Prices Aren't Available," *Journal of Economic Perspectives* 26(4), Fall 2012, 27-42.

Carson, Richard T., Nicholas E. Flores, Kerry M. Martin, Jennifer L. Wright, "Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods," *Land Economics* 72(1), February 1996, 80-99.

Carson, Richard, and Theodore Groves, "Incentive and Informational Properties of Preference Questions," *Environmental Resource Economics* 37, 2007, 181-210.

Carson, Richard, Theodore Groves, and John A. List, "Toward an Understanding of Valuing Non-Market Goods and Services," mimeo, UCSD, 2011.

Carson, Richard, and W. Michael Hanemann, "Contingent Valuation," *Handbook of Environmental Economics*, Volume 2, K.-G. Maler and J.R. Vincent, eds., Elsevier, 821-936.

Champ, P. A., R. C. Bishop, T. C. Brown, and D. W. McCollum, "Using donation mechanisms to value nonuse benefits from public goods," *J Environ Econ Manage* 33, 1997, 151–162.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz, "The Short- and Long-Term Effects of Measuring Intent to Repurchase," *Journal of Consumer Research* 31, 2004, 566-572.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz, "Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research," *Journal of Marketing* 69, April 2005, 1-14.

Cummings, R. G., and L. O. Taylor, "Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method," *American Economic Review* 89(3), 1999, 649–665.

Fox, John A., Jason F. Shogren, Dermot J. Hayes, and James B. Kliebenstein, "CVM-X: Calibrating Contingent Values with Experimental Auction Markets," *American Journal of Agricultural Economics* 80(3), August 1998, 455-465.

Harrison, G., R. Beekman, L. Brown, L. Clements, T. Mc Daniel, S. Odom, and M. Williams, "Environmental damage assessment with hypothetical surveys: The calibration approach," in M. Bowman, R. Brannlund, and B. Kristroem (eds.), *Topics in Environmental Economics*, Kluwer, Amsterdam, 1997.

Infosino, William J., "Forecasting New Product Sales from Likelihood of Purchase Ratings," *Marketing Science* 5(4), Special Issue on Consumer Choice Models, Autumn, 1986, 372-384.

Jackman, Simon, "Correcting surveys for non-response and measurement error using auxiliary information," *Electoral Studies* 18, 1999, 7-27.

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchinix, and Jason F. Shogren, "Preference elicitation under oath," mimeo, November 2010.

Jamieson, Linda F., and Frank M. Bass, "Adjusting Stated Intention Measures to Predict Trial Purchase of New Products: A Comparison of Models and Methods," *Journal of Marketing Research*, 26(3), August 1989, 336-345.

Johansson-Stenman, O. and H. Svedsäter, "Choice experiments and self image: Hypothetical and actual willingness to pay," Working Paper, Gothenburg University, 2003.

Juster, T., *Anticipations and Purchases*, Princeton, NJ: Princeton University Press, 1964.

Kang, Min, Antonio Rangel, Mikael Camus, and Colin F. Camerer. "Hypothetical and real choice differentially activate common valuation areas," *Journal of Neuroscience*, 2011, 31: 461-468.

Katz, Jonathan N., and Gabriel Katz, "Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout," *American Journal of Political Science* 54(3), July 2010, 815–835.

Kraut, Robert E., and John B. McConahay, "How Being interviewed Affects Voting: An Experiment," *Public Opinion Quarterly* 37(3), Autumn 1973, 398-406.

Krueger, Alan B., and Ilyana Kuziemko, "The demand for health insurance among uninsured Americans: results of a survey experiment and implications for policy," *Journal of Health Economics* 32(5), 2013, 780-793.

Kurz, Mordecai, "Experimental approach to the determination of the demand for public goods," *Journal of Public Economics* 3, 1974, 329-348.

Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella, "Penalized Regression, Standard Errors, and Bayesian Lassos," Bayesian Analysis 5(2), 2010, 369-412.

Leeb, Hannes, and Bekedikt M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, 2005, 21-59.

Levy, I., S. C. Lazzaro, R. B. Rutledge, and P. W. Glimcher, "Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing," *Journal of Neuroscience* 31, 2011, 118-125.

List, John A., and Craig A. Gallet, "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? Evidence from a Meta-Analysis," *Environmental and Resource Economics* 20, 2001, 241–254.

List, John A., and Jason F. Shogren, "Calibration of the difference between actual and hypothetical valuations in a field experiment," *Journal of Economic Behavior and Organization* 37, 1998, 193-205.

List, John A., and Jason F. Shogren, "Calibration of Willingness-to-Accept," *Journal of Environmental Economics and Management* 43, 2002, 219-233.

Little, Joseph and Robert Berrens, "Explaining Disparities between Actual and Hypothetical Stated Values: Further Investigation Using Meta-Analysis," *Economics Bulletin* 3(6), 2004, 1−13

Loomis, J., K. Traynor, and T. Brown, "Trichotomous choice: a possible solution to dual response objectives in dichotomous choice contingent valuation questions," *J Agric Resour Econ* 24(2), 1999, 572–583.

Louviere, J., "Conjoint analysis," in R. Bagozzi (ed.), *Advanced Methods in Marketing Research*, Cambridge: Blackwell Business, 1993.

Louviere, J., D. Hensher, and J. Swait, *Stated Choice Methods: Analysis and Application*, Cambridge: Cambridge University Press, 2000.

Mansfield, Carol, "A Consistent Method for Calibrating Contingent Value Survey Data," *Southern Economic Journal*, 64(3), January 1998, 665-681.

Morrison, D., "Purchase Intentions and Purchase Behavior," *Journal of Marketing* 43, 1979, 65-74

Morrison, Mark, and Thomas C. Brown, "Testing the Effectiveness of Certainty Scales, Cheap Talk, and Dissonance-Minimization in Reducing Hypothetical Bias in Contingent Valuation Studies," *Environmental Resource Economics* 44, 2009, 307–326.

Morwitz, Vicki G., Joel H. Steckel, Alok Gupta, "When do purchase intentions predict sales?" *International Journal of Forecasting* 23, 2007, 347–364.

Murphy, Janes J., P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weatherhead, "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation," *Environmental and Resource Economics* 30, 2005, 313–325.

National Oceanic and Atmospheric Association, "Natural Resource Damage Assessments: Proposed Rules," *Federal Register* 59, January 1994, 1142.

Polak, J., and P. Jones, "Using stated-preference methods to examine travelers preferences and responses," in P. Stopher and M. Lee-Gosselin (eds.), *Understanding Travel Behavior in an Era of Change*, Oxford: Pergamon, 1997.

Rothschild, David, "Forecasting elections: comparing prediction markets, polls, and their biases," *Public Opinion Quarterly* 73(5) 2009, 895–916.

Rothschild, David, and Justin Wolfers, "Forecasting elections: voter intentions versus expectation," mimeo, 2011.

Saez, Emmanuel, "Details Matter: The Impact of Presentation of Information on the Take-up of Financial Incentives for Retirement Saving," *American Economic Journal: Economic Policy* 1(1), February 2009, 204-228.

Shogren, J., "Experimental Markets and Environmental Policy," *Agr. Res. Econ. Rev.* 3, October 1993, 117-29.

Shogren, Jason, "Experimental Methods and Valuation," in K.-G. Mäler and J.R. Vincent (eds.), *Handbook of Environmental Economics, Volume 2*, Elsevier, 2005, 969-1027.

Shogren, Jason F. "Valuation in the lab," *Environmental & Resource Economics* 34, 2006, 163–172.

Smith, Alec, B. Douglas Bernheim, Colin F. Camerer, and Antonio Rangel, "Neural activity reveals preferences without choice," *American Economics Journal: Microeconomics* 6(2), 2014, 1-36.

Tibshirani, Robert, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Societ,. Series B (Methodological)* 58(1), 1996, 267-88.

Tusche, A., S. Bode, and J. D. Haynes, "Neural responses to unattended products predict later consumer choices," *Journal of Neuroscience* 30, 2010, 8024-8031.

West, Kenneth D., "Forecast Evaluation, " in G. Elliott, C.W.J. Granger, and A.G. Timmermann (eds.), *Handbook of Economic Forecasting*, Amsterdam: North Holland, Vol. 1, Ch. 3, 2006, 99-134.

White, Halbert, "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review* 21(1), February 1980, 149-170.

Yates, J. Frank, "External correspondence: Decompositions of the mean probability score," *Organizational Behavior and Human Performance* 30, 1982, 132-156.

Table 1: Predictive accuracy for benchmark methods

Benchmark method	**Predicting from $0.75 to $0.25**				**Predicting from $0.25 to $0.75**			
	Normalized Avg Effect	MSPE	Calibration (level)	Calibration (change)	Normalized Avg Effect	MSPE	Calibration (level)	Calibration (Δ)
Methods using limited choice data								
Myopic	0.000[a]	93.3	1.001	NA	0.000[a]	93.3	0.690[a]	NA
Structural	2.544[a]	239.6	0.844[c]	0.033[a]	2.909[a]	281.7	2.638[b]	0.322[a]
HPF, Difference method	1.987[a]	137.4	0.610[a]	0.248[a]	1.987[a]	137.4	0.500[a]	0.248[a]
HPF, Levels method	2.409[a]	243.1	0.474[a]	0.207[a]	0.577[a]	103.4	0.466[a]	0.240[a]
Methods using additional choice data								
Average Δ	1.000	37.4	1.003	NA	1.000	37.4	0.688[a]	NA
RPF predictor	0.998	37.9	0.996	-3.984[a]	0.998	26.2	0.992	0.993
Augmented predictors, OLS	1.008	39.4	0.971	0.315[a]	1.027	35.9	0.805	0.529[b]
Augmented predictors, LASSO	0.999	36.9	1.023	0.558[d]	0.998	25.5	1.042	0.903

Shaded rows use real purchase frequencies not only at the starting price for all items, but also at the alternative price for 80% of items (not including the item of interest). Superscripts indicate that the ideal value (unity) lies outside a given percentile interval for the bootstrap distribution, as follows: for *a*, 0.5^{th}-to-99.5^{th}; for *b*, 2.5^{th}-to-97.5^{th}; for *c*, 5^{th}-to-95^{th}; for *d*, 10^{th}-to-90^{th}.

Table 2: Predictive accuracy of optimized specifications

Model	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75			
	Normalized Avg Effect	MSPE	Calibration (level)	Calibration (change)	Normalized Avg Effect	MSPE	Calibration (level)	Calibration (Δ)
Best-performing benchmark: Augmented predictors, Lasso	0.999	36.9	1.023	0.558[d]	0.998	25.5	1.042	0.903
All hyp. & ratings								
LASSO	0.850	30.6	0.948	0.932	1.083	29.2	0.759[b]	0.847
OLS – CV-Calib optimized	1.141	31.5	0.908	0.738[d]	0.922	27.7	0.772[b]	0.993
OLS – CV-MSPE optimized	1.014	30.6	0.892	0.694[b]	1.125	31.6	0.754[b]	0.678[c]
OLS – CV-AMPE optimized	1.148	33.5	0.902	0.669[c]	0.943	27.9	0.771[a]	0.950
Hyp. only								
Univariate	0.927	36.2	0.919[b]	0.531[a]	0.942	36.4	0.692[a]	0.523[a]
Bivariate	1.022	32.3	0.924[c]	0.645[a]	0.975	31.8	0.728[a]	0.674[a]
All hyp., LASSO	1.276[d]	34.1	0.938	0.773[d]	1.547[c]	47.4	0.752[b]	0.681[c]
All hyp., ratings, & phys.								
LASSO	0.887	31.4	0.973	0.868	1.030	27.5	0.773[a]	0.924
LASSO, with RPF	1.018	29.6	1.052	0.963	1.005	22.5	1.050	0.897

Shaded rows use real purchase frequencies not only at the starting price for all items, but also at the alternative price for 80% of items (not including the item of interest). Superscripts indicate that the ideal value (unity) lies outside a given percentile interval for the bootstrap distribution, as follows: for *a*, 0.5^{th}-to-99.5^{th}; for *b*, 2.5^{th}-to-97.5^{th}; for *c*, 5^{th}-to-95^{th}; for *d*, 10^{th}-to-90^{th}.

Figure 1: The relationship between real and hypothetical purchase frequencies.