NBER WORKING PAPER SERIES

DO HYPOTHETICAL CHOICES AND NON-CHOICE RATINGS REVEAL PREFERENCES?

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
July 2013

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, and UCSD for helpful comments. Detailed suggestions from Richard Carson and Laura Taylor were especially helpful. We are also grateful to Irina Weisbrott 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.

Do Hypothetical Choices and Non-Choice Ratings Reveal Preferences?
B. Douglas Bernheim, Daniel Bjorkegren, Jeffrey Naecker, and Antonio Rangel
NBER Working Paper No. 19269
July 2013
JEL No. C91,D12,H31,Q51

ABSTRACT

We develop a method for determining likely responses to a change in some economic condition (e.g., a policy) for settings in which either similar changes have not been observed, or it is challenging to identify observable exogenous causes of past changes. The method involves estimating statistical relationships across decision problems between choice frequencies and variables measuring non-choice reactions, and using those relationships along with additional non-choice data to predict choice frequencies under the envisioned conditions. In an experimental setting, we demonstrate that this method yields accurate measures of behavioral responses, while more standard methods are either inapplicable or highly inaccurate.

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

Daniel Bjorkegren
Harvard University
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

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

1. Introduction

A central problem in microeconomics is to predict the distribution of households' choices in not-yet-observed situations (e.g., after some policy intervention). With few exceptions, mainstream economists attack that problem using one of two approaches. The first is to extrapolate behavior from closely related variation in the economic environment using reduced-form models. A canonical example is the estimation of a demand curve from a collection of price-quantity observations with exogenous price variation. The second approach is to extrapolate behavior from more distantly related variation in the economic environment using structural models; in effect, the assumed structure links the observed variation to the intervention of interest. A canonical example involves the prediction of choices from piecewise-linear budget sets in a context where all observed choices are made from linear budget sets.

While these approaches serve economists well in a wide variety of contexts, they also have limitations. Often, it is impossible to identify exogenous variation (or in some cases *any* variation) in the economic environment that is closely related to the intervention of interest, in which case the first approach is unavailable. Moreover, when the observed variation in conditions is only distantly related to the intervention of interest, the second approach can require strong and potentially implausible assumptions. These problems are particularly acute in behavioral economics, due to the well-documented importance of context and framing effects (see, e.g., Camerer et al., 2004, Bertrand et al., 2005, Saez, 2009). When the intervention of interest involves a novel decision frame, the first approach is not available, while the second requires a deeper structural understanding of the psychological processes that generate framing effects than, in most cases, we currently possess.

On rare occasions, typically when the problems with the two conventional approaches are thought to be especially severe, economists turn (often reluctantly) to a third approach: ask people, hypothetically, what they would choose in the settings of interest. 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). Unfortunately, there is little if any basis in the literature for judging which estimates more accurately reflect the true elasticities of the uninsured.

The use of hypothetical questions has been explored in a sizable literature on stated preference (SP) techniques and the contingent valuation method (CVM); for reviews, see Shogren (2005, 2006), Carson and Hanemann (2005), and Carson (2012). It is well-established that answers to standard hypothetical questions are systematically biased. Two classes of solutions have been examined: one attempts to "fix" the hypothetical question; the other seeks to correct for the bias through *ex post* statistical calibration. We elaborate on both approaches in Section 2. Their limitations are widely acknowledged, and consequently their use is largely confined to contexts where choice data pertaining to closely related decisions are *entirely* unavailable (e.g., in the environmental context, to value non-market goods such as pristine coastlines, biodiversity, and the like),[2] rather than merely imperfect.

Despite the limitations of stated preference techniques, measures of elicited preferences are indisputably correlated with actual choices, and are therefore potentially useful, if not as *predictions* of real choices, then at least as *predictors*. Furthermore, once we stop interpreting hypothetical choices literally and instead treat them merely as variables containing predictively useful information, additional possibilities open up. Other potential predictors include any reaction to elements of a contemplated opportunity set that occur when an individual is *not* engaged in actual decision-making. These *non-choice reactions* could include ratings or even physiological reactions assessed while an individual contemplates a hypothetical choice or consumption experience.

These observations suggest a more general strategy for predicting choices in as-yet unobserved economic environments: elicit non-choice reactions to a range of choice environments (some observed, others unobserved), uncover statistical relationships between real choices and combinations of non-choice reactions that are stable over reasonably broad domains, and use those relationships to predict behavior out of sample for the unobserved environments. 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).

While our strategy bears some relation to the calibration approaches explored in the SP literature (mentioned above and discussed in Section 2), it has several distinctive features.

First, instead of treating the individual as the unit of observation and predicting each choice, we treat the decision problem at the unit of observation and predict choice distributions. Accurate prediction of each individual's choice is not only potentially more difficult, but also

[2] 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).

unnecessary in most economic applications. Notably, to predict the distribution of choices made by a fixed population of individuals in an as-yet unobserved decision problem, one needs to account for the differences between decision problems, rather than the differences between individuals within that population.

Second, we construct predictors based on distributions of various types of non-choice reactions, including hypothetical versions of the same decision problems, as well as other types of ratings. We focus mainly on non-choice reactions that are *portable* in the sense that they are likely to have context-independent implications over broad domains. For example, the implication of the statement that an individual would choose A over B, or likes A better than B, does not depend to any great extent on what A and B represent. Consequently, these reactions are at least arguably portable from one context to another. In contrast, the implication of the statement that A tastes better than B differs dramatically depending on whether A and B are, say, food items or articles of clothing. We focus mainly on portable non-choice responses because the relationships between real choices and non-choice responses with context-specific interpretations are likely to be stable only over narrow domains, which would limit their usefulness.

Third, we use the aforementioned predictors in combination. Thus, instead of adopting a single elicitation protocol for hypothetical choices, we allow for the possibility that different protocols may succeed in capturing somewhat different information about real decisions, and use the elicited variables as co-predictors. Also, because the degree of hypothetical bias differs from one context to another for systematic reasons, non-choice variables that are correlated with factors that contribute to the degree of hypothetical bias make valuable co-predictors.

Fourth, our focus is on out-of-sample predictive performance – in particular, on the ability to forecast choice distributions in as-yet unobserved economic environments. Because we envision using these methods in settings where (imperfect) choice data are also available, we run "horse races" between choice-based and non-choice-based methods.

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 either \$0.25 or \$0.75, to be consumed during a waiting period, and collect data on purchase frequencies for many items at both prices. We then set the following task: supposing that one only observes purchase frequencies at a single 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 the *other* price? 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). Instead, a conventional economist might proceed by building a structural model, 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. We estimate such models, invoking apparently reasonable assumptions. However, when we use them to predict purchase frequencies out of sample at the alternative price, they perform poorly – indeed, worse than the myopic prediction of no change in purchase frequencies.

An alternative is to use the SP approach; i.e., ask people what they would choose at the alternative price, and take those responses as indicating the actual purchase frequency. However, this approach also yields poor predictions, likewise underperforming the myopic benchmark. Although several alternative elicitation protocols appear to reduce the overall degree of hypothetical bias (consistent with the SP literature), they do not generally improve the quality of predictions in this setting according to a variety of other metrics. Indeed, in most instances they also underperform the myopic benchmark.

The final (and most critical) step in our analysis is to evaluate the performance of our alternative approach. We estimate statistical relationships between real purchase frequencies and non-choice reactions at the price that is assumed to have prevailed (e.g., $0.25), and use those relationships along with additional data on non-choice reactions to predict real purchase frequencies at the alternative price (e.g., $0.75). The specifications favored by within-sample model selection criteria predict purchase frequencies out of sample at the alternative price with a high degree of accuracy; e.g., in the best such specifications, the average predicted change in demand is within a few percent of the average actual change. Moreover, the performance of this approach roughly matches 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.

2. Related Literature

Despite the well-recognized limitations of choice data, standard economics makes little use of non-choice alternatives. Here we identify the main exceptions and explain their relationships to our approach. We also discuss related literature from other disciplines.

There is a voluminous literature on stated preference (SP) techniques, particularly in the context of 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; in contrast, we explore the use of non-choice data as an alternative to choice data even when the latter are available (but are not ideal).

It is well-established that answers to standard hypothetical questions are systematically biased, typically in the direction of overstating willingness-to-pay (WTP) and toward alternatives that are viewed as more "virtuous."[3] Two classes of solutions have been examined. One is to "fix" the hypothetical question through 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). Our approach is more closely related to a second class of solutions, involving *ex post* statistical calibration techniques (in particular, Shogren, 1993, Blackburn, Harrison, and Rutstrom, 1994, Fox et al., 1998, List and Shogren 1998, 2002, and, to a lesser extent, Mansfield, 1998). *Ex post* calibration (which can be traced to Kurz, 1974, and was considered by National Oceanographic and Atmospheric Association, 1994) exploits a statistical relationship between real and hypothetical choices and, like our approach, treats 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 critical respects. The most important differences are related to the fact that all of the calibration studies listed above are concerned with predicting individual-specific choices for a single decision problem at a time, rather than choice distributions for as-yet unobserved decision problems. Thus, they account for differences in bias across individuals (for a given decision problem) that are related to socioeconomic and demographic characteristics, but they do not account for (and cannot predict) differences in hypothetical bias across choice problems (for a stable population). On the contrary, List and Shogren (1998, 2002) emphasize that hypothetical bias is good- and context-specific, so that individual-level calibration does not reliably transfer from one setting to another.[4] Yet psychological studies also suggest that hypothetical bias is

[3] 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).

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

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 drawn from the same target population. In contrast, in individual-level calibration studies, subjects make real choices after making hypothetical ones, which introduces the possibility that the former contaminate the latter.[5] 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), 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,[6] 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 (2012) focus specifically on passive non-choice neural reactions, and provide proof-of-concept that those types of reactions predict choices.[7] 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 Bass, 1989). Its relationship to our work is similar to that of the SP/CVM literature on *ex post*

[5] 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).

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

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

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 (see Louviere, 1993, Polak and Jones ,1993, 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.[8] 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.

3. Experimental procedures and data

Our analysis employs the following data: (1) real choice frequencies for a large number of items at two different prices, (2) hypothetical choice frequencies, elicited through various protocols, for the same items and prices, and (3) response frequencies for questions eliciting other ratings of the same items (with price a factor in some but not all questions). We chose the ratings questions with the object of obtaining responses that might contain information about the size of the gap between real and hypothetical choice frequencies for a given item, on the grounds that such responses would likely make good co-predictors. To minimize cross-contamination of responses, we used multiple non-overlapping subject groups, as described below.

We conducted the experiment at the Stanford Economic Research Laboratory (SERL). At the outset of each 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 described below). Sessions took place in mid-afternoon, when subjects are typically hungry.

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

All subjects performed first-stage tasks pertaining to a fixed set 189 snack food items. 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).

A total of 365 subjects participated in the experiment (181 males, 184 females). Each subject was paid a participation fee between $20 and $30.[9] Subjects were divided into multiple treatment groups, with each subject participating in a single treatment (except as noted below). The nature of the treatments and the sizes of the various groups were as follows. In all cases the stimuli (food items, or item-price pairs) were presented in random order.

Treatment R (30 subjects): Subjects made real choices using the strategy method. Each item appeared twice, once with a price of 25 cents and once with a price of 75 cents. In each case, the subject had to decide whether to buy the item at the specified price. The subject was told that, prior to stage 2 of the experiment, one choice problem would be selected at random and implemented, with all equally likely. Any subject who opted to make a purchase in the selected choice problem paid the indicated price out of the participation fee, and was given the item as a snack during the waiting period. Any subject who opted not to make a purchase in the selected choice problem received no snack and retained the entire participation fee.

Treatments H and HD (28 subjects each): Subjects made the same choices as in treatment R, but were aware that all of their decisions were hypothetical, and would not be implemented. There is no difference between these two treatments; the "D" in "HD" stands for "duplicate." Duplicating treatment H allows us to investigate whether it is better to use additional subjects to increase sample sizes or answer new questions.

Treatment M (35 subjects)*:* Subjects made the same choices as in treatment R, but were told in advance that all but five would be hypothetical. The five real choices were interspersed among the hypothetical choices, but clearly indicated when they were presented. For each subject, the five items were drawn at random from a larger group of fifteen, selected for their representativeness,[10] and each was offered at a price of 75 cents. The purpose of this "mixed" treatment is to investigate the concern, discussed below, that the low probability with which any given choice problem was implemented in treatment R influenced purchase frequencies (possibly by inducing subjects to treat the "real" choices as hypothetical).

[9] We adjusted the fee upward when the response rate to our subject solicitation was low, and downward when it was high.

[10] Specifically, the distribution of purchase frequencies (among Group R) for the 15 items mirrors the distribution of purchase frequencies for all 189 items.

Treatment HCT (28 subjects): Subjects performed that same task as in treatment H, but a "cheap talk" script (as in Cummings and Taylor, 1999) was added to the experimental instructions, with the objective of inducing subjects to take the hypothetical choices more seriously, and thereby minimize hypothetical bias.[11]

Treatment HL (28 subjects): Subjects performed the same task as in treatment H, but the questions were modified to elicit the likelihood that the subject would buy the item using a five-point scale (1="very likely," 3="uncertain," 5="very unlikely"), rather than a Yes/No decision. The object of this treatment is to collect information that permits us to distinguish between statements about which subjects are reasonably certain, and those about which they are uncertain, analogously to Champ et al. (1997).

Treatment HV (28 subjects): Subjects performed the same task as in treatment HL, except they were asked to indicate how they thought a typical undergraduate of their own gender would answer. The object of these "vicarious" questions is to eliminate image concerns and hence elicit more honest answers, analogously to Rothschild and Wolfers (2011).

Treatment HWTP (28 subjects): Subjects expressed a hypothetical willingness to pay (WTP) for all of the food items, each of which appeared only once. We employed this protocol because much of the literature explores the accuracy of hypothetical WTPs rather than binary choices. We used the same subjects for treatments HWTP and L (below).[12]

Treatment SWB (28 subjects): For each potential outcome, subjects indicated their anticipated subjective well-being: "How happy would you be if you received this item (and ONLY this item) to eat as a snack during the second part of this experiment, and a price of $X was deducted from your show-up payment?" (with 1="very unhappy" and 7="very happy"). Each item appeared twice, once with a price of 25 cents and once with a price of 75 cents.

Treatment N (28 subjects): Subjects indicated whether each potential outcome would elicit social approval or disapproval: "Imagine that a subject in this experiment paid X cents to eat the item as a snack during the second part of the experiment. Would the typical person approve or disapprove of this purchase?" (with 1="strong disapproval" and 7="strong approval"). These ratings are intended to capture social norms and image concerns.

Treatment L (28 subjects): Subjects provided liking ratings for each item: "How much would you like to eat this item during the second part of the experiment?" (with 1="not at all"

[11] 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).
[12] We combined treatments HWTP and L because each required subjects to make fewer responses (i.e., one response for each item, rather than two as in treatment R and other hypothetical choice treatments).

and 7="very much"). We include this treatment because liking ratings are known to be correlated with choices. As noted above, we used the same subjects for treatments L and HWTP.

Treatment S (29-38 subjects):[13] Subjects answered some or all of the following additional questions concerning the food items (answers scaled 1-5): 1) "How much would you later regret eating this snack?" 2) "How tempting is this item?" 3) "If you had no concerns about diet or health, how much would you enjoy eating this item?" 4) "Is this item generally good or generally bad for you?" 5) "Would others form a positive or negative impression of you if they saw you eating this snack?" 6) "Are people likely to understate or overstate their inclination to pick this snack?" The responses to these questions may be useful for predicting choices because each question potentially gets at factors related to the degree of hypothetical bias. Questions 1 through 4 address the degree to which immediate gratification conflicts with longer term considerations: we conjectured that hypothetical choices will be more sensitive to long-term costs, and less sensitive to immediate gratification, than real choices. Question 5 addresses concerns for social image: we conjectured that hypothetical choices will be more sensitive to image concerns than real choices. Finally, the purpose of question 6 is to determine whether subjects can provide subjective assessments of hypothetical bias that would be useful for the purpose of predicting choices, even if the sources of the bias remain unclear.

The Appendix contains the instructions provided to members of each treatment group, including a screenshot for a representative question (Figure A.1). The experimental protocol was reviewed and approved by Stanford University's IRB.

One potential concern is that members of treatment group R may have viewed the "real" choices as hypothetical in light of the low implementation probabilities (one in 378). That possibility is strongly refuted by the results presented in the next section. In particular, purchase frequencies are on average significantly higher for treatment H than for treatment R (consistent with the general finding in the literature concerning the direction of hypothetical bias); the cross-choice-task variance of the purchase frequency is considerably higher for treatment H than for treatment R; the average price sensitivity implied by the purchase frequencies is much larger for treatment H than for treatment R; and the variance of the implied price response is much larger

[13] We collected 29 subject responses to questions 1, 5, and 6, and either 38 or 31 subject responses (depending on the item) to questions 2, 3, and 4.The variation in sample sizes across items for questions 2, 3, and 4, which occurred because of the manner in which the experiment evolved, is not ideal, but we doubt it has a meaningful impact on our results. Initially we collected responses to questions 1, 5, and 6 from a group of 9 subjects, and responses to questions 2, 3, and 4 from a group of 16 subjects, but concerning only 120 of the 189 items. We then collected responses to questions 1, 5, and 6 from a group of 20 subjects, and responses to questions 2, 3, and 4 from a group of 22 subjects, concerning all 189 items. We then collected responses to all six questions from a group of 9 subjects, but only for the 69 items for which we collected no data from the first two groups.

for treatment H than for treatment R. Plainly, despite the low implementation probabilities, subjects treated the real and hypothetical questions much differently.

It does not follow, however, that treatment R subjects viewed their choices as entirely real; they may have adopted a hybrid outlook, part real and part hypothetical. To evaluate that possibility, we employed data from treatment M. Within that group, the implementation probability for each real choice was 20 percent rather than 1/378. In total, we elicited 175 real choices through treatment M, pertaining to 15 distinct items. We then pooled that data with the 450 choices involving the same 15 items (at a price of $0.75) from treatment R, and estimated a probit regression relating the purchase decision to a set of 15 product dummies as well as a treatment M dummy. If the treatment R subjects viewed their choices as real, the coefficient for the treatment M dummy should be zero; if instead they treated those choices as partially hypothetical, then the treatment M 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.[14] 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 treatment M than for treatment R, which is inconsistent with the hypothesis that participants in treatment R were more inclined to view their "real" choices as hypothetical than were participants in treatment M.

We are not surprised by the finding that participants in treatment R viewed their choices as real. After all, they had as much at stake as someone making a single purchase decision, 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.[15]

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

[14] For treatment M, 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 treatment R). Thus, the presence of real choices in treatment M did not induce subjects to treat their hypothetical choices as real; they still suffered from hypothetical bias.

[15] It is important to acknowledge, however, that the pertinent studies involved far fewer contingent choices than in our treatment R.

(either \$0.25 or \$0.75), at which actual purchases have been observed. There is a proposal to change all of 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 somewhat artificial 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 it is endogenous and no valid instrument is available. For instance, the objective may be to gauge responses to a proposed policy change that has no close precedent.

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 does respond to price: the RPFs average 27.76% for a price of \$0.25 and 20.26% for a price of \$0.75 (p = 0.001). 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 modest size of treatment group R (30 subjects), some of that variation reflects sampling error.

To determine whether sampling variation could account for the observed variance of the RPF across items for a fixed price, we perform the following calculation. The reported sample variance of the RPF across items at a fixed price P_j is $s_{Rj}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_{ij} - \bar{R}_j\right)^2$, where R_{ij} represents the RPF for item i when it sells for price P_j, $\bar{R}_j$ represents the overall average RPF at price P_j, and N denotes the number of items. Treating both the selection of items and the choice of subjects as random, and allowing for the possibility that the choices of a randomly selected subject may be correlated across decisions, one can show that

$$E\left[s_{Rj}^2\right] = \sigma_{Rj}^2 + \sigma_{\omega j}^2(1 - \rho_{Rj}),$$

where σ_{Rj}^2 denotes the true variance of the population RPF across items (given the distribution from which the items are selected), $\sigma_{\omega j}^2$ denotes the variance of the sampling error $\omega_{ij} = R_{ij} - R_{ij}^P$ across items, and ρ_{Rj} is the correlation between the sampling errors of two randomly selected items. For any given value of R_{ij}^P, the distribution of the sampling error is binomial, with $Var\left(\omega_{ij}\middle|R_{ij}^P\right) = R_{ij}^P(1 - R_{ij}^P)/N$, where *N* is the group size. Noting that the preceding formula is concave in R_{ij}^P, we have $\sigma_{\omega j}^2 < \bar{R}_j^P(1 - \bar{R}_j^P)/N$ (where $\bar{R}_j^P$ is the mean of R_{ij}^P for all of the items). Using $\bar{R}_j$ as an estimate of $\bar{R}_j^P$, we conclude that $\sigma_{\omega,0.25}^2 < \frac{27.76(100-27.76)}{30} = 66.8$ and $\sigma_{\omega,0.75}^2 < \frac{20.26(100-20.26)}{30} = 53.9$. In addition, the correlation between sampling errors across item-price pairs is likely positive (e.g., because hungry subjects are more inclined to buy all items). Therefore, it is reasonable to infer that $\sigma_{\omega j}^2(1 - \rho_j)$ is even smaller. In contrast, $s_{R,0.25}^2 = 120.7$ and $s_{R,0.75}^2 = 83.2$. We conclude that at least 40% of the variance in the measured RPFs at either price – and likely much more – reflects real variation in the appeal of the items pairs.

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 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 7), 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 (or mean prediction error, MPE), mean-squared prediction error (MSPE), and calibration.[16]

[16] 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 2, "calibration" has an entirely different meaning in the literature on SP techniques.

To be more specific, let $\hat{R}_{i2}$ denote the predicted RPF for item *i* at price P_2. We compute the overall bias for the level of predicted demand, $\hat{R}_{i2}$, as follows: MPE $= \frac{1}{N}\sum_{i=1}^{N}(\hat{R}_{i2} - R_{i2})$. Similarly, we compute the MSPE for the level of predicted demand as follows: MSPE $= \frac{1}{N}\sum_{i=1}^{N}(\hat{R}_{i2} - R_{i2})^2$. The MPE and the MSPE for the predicted change in demand, $\hat{R}_{i2} - R_{i1}$, are also of interest, but they are mathematically identical to the MPE and the MSPE for the level of demand, $\hat{R}_{i2}$. Consequently, in what follows we will simply refer to the MPE and 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:

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

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 MPE 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. In practice, the task of achieving good calibration is typically more challenging for predicted changes in demand than for predicted levels.

5. Benchmarks

When evaluating the quality of a prediction, it is important to have in mind benchmarks that help one gauge what constitutes "good" performance. In this section, we present two classes of benchmarks. The first involves prediction methods that employ no more data on actual choices than the methods we wish to evaluate. The second class involves methods that use more

extensive data on actual choices (and hence would not be feasible if choice data were limited in the way our prediction task assumes).

A. Benchmarks that use limited choice data

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.

For additional benchmarks we employ standard econometric tools. Reduced-form methods are not applicable because the available choice 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 the quantity of a serving across items (controlling for the items' characteristics), because a difference in quantity implies a difference in cost per gram.

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 $V_i = v_i q_i$, where v_i is the value of the item per gram and q_i is the number of grams. 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; if the expression for V_i included a constant term, the method we use below for forecasting purchase decisions at the alternative price P_2 would not work.

In principle, one could treat each v_i as a free parameter, but that would make unreasonable demands of our data. Instead we assume that $v_i = X_i\beta$, where X_i is a vector of characteristics and β is a vector of parameters. The vector X_i can include a constant, binary variables for various food categories, as well as continuous variables capturing nutritional content, where nutrients are measured per gram (because v_i is value per gram).

With a little algebra, one can reformulate this model in latent variable form. Specifically, we define

$$Y_{is}^{*} \equiv \delta + Z_i\gamma + \eta_{is}\ , \qquad (2)$$

where $\delta = -\frac{P_1}{\sigma}$, $Z_i = X_i q_i$ (so that Z_i potentially includes grams per serving, category dummies interacted with grams per serving, and nutrients measured per serving rather than per gram), and $\gamma = \frac{\beta}{\sigma}$. The condition $V_i + \varepsilon_{is} \geq P_1$, which governs the purchase decision, is then equivalent to $Y_{is}^{*} \geq 0$. Thus, we can estimate (2) as a logistic regression.

Once we estimate this model with choices made at the price P_1, we can use it to forecast choices at some alternative price, P_2. At the new price, the value of the latent variable becomes

$$\hat{Y}_{is}^{*} = \delta\left(\frac{P_2}{P_1}\right) + Z_i\gamma + \eta_{is}\ .$$

So, for example, if P_1 is \$0.25 and P_2 is \$0.75, we simply multiply the constant in the estimated equation by 3 (which should reduce $\hat{Y}_{is}^{*}$ because we expect δ to be negative), compute the implied probability that subject s will purchase item i based on the estimated distribution of η_{is}, and then average those probabilities over subjects to obtain the predicted RPF for item i.

We implement four versions of the preceding model, which include different variables in the vector Z_i. For the simplest, Z_i includes only grams per serving. Additional variants add variables measuring nutrients per serving, category dummies interacted with grams per serving, or both. We omit the model estimates for the sake of brevity; generally, they seem reasonable.

As shown in the second through fifth lines of Table 1, all of these models perform terribly out of sample. The overall bias for the predicted price response is roughly twice the actual price response; thus, the models all imply price responses roughly three times as large as the actual responses. They also perform 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 from \$0.25 choices to \$0.75 choices, and calibration for changes is quite poor (particular when predicting from \$0.75 choices to \$0.25 choices).

We emphasize that these results should not be interpreted as a general indictment of structural methods. Rather, they show that the prediction task we have set ourselves is a challenging one. That is why the success of our method, documented below, is notable.

B. Benchmarks that also use additional choice data

For our second set of benchmarks, we assume that the set of observed choices includes purchase decisions for the item of interest at the price P_1, plus decisions 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).

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; the others include variables that identify food categories and measure nutritional context. We use OLS because it has desirable forecasting properties (see, e.g., White, 1980). However, recognizing that OLS is susceptible to the overfitting problem in contexts where the number of potential predictors is large relative to the number of observations, we also employ LASSO (the Least Absolute Shrinkage and Selection Operator, due to Tibshirani, 1996).

Measures of predictive performance appear in the lower half of Table 1. All of these approaches yield substantial improvements over the myopic benchmark. Much of the gain is achieved simply by assuming that the price response for each item would be the same as the average response for other items. 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 to \$0.75 choices, but not when predicting the other way around. Relative to that benchmark, 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.

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. The accuracy of hypothetical responses as predictions

We begin by evaluating the accuracy of hypothetical responses, treating them as predictions rather than as predictors. Henceforth we abbreviate "hypothetical purchase frequency" as HPF. The RPF always refers to purchase frequencies for treatment R, but the HPF can refer to treatments H, HCT, HL, HV, or HWTP (and the treatment is indicated).

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). As we will see, this is typically the case.

Table 2 contains summary statistics for our various hypothetical choice frequencies (pooling data from all 378 item-price pairs), as well as measures of predictive accuracy for the difference method. With that method, the absolute value of the average bias is the same when predicting from \$0.75 choices to \$0.25 choices and the other way around; only the sign changes. The table adopts the convention of reporting the average bias for the \$0.75-to-\$0.25 predictions. Likewise, the MSPE is exactly the same in either case, as is calibration for the change in RPF. However, calibration for levels differs according to whether we are predicting \$0.75 or \$0.25 choices, so we report both. Table 3 reports those same measures for the levels method. For purposes of comparison, Table 2 also reports overall summary statistics for the RPF, and both tables include the performance statistics for myopic predictions. The remainder of the table evaluates the predictive accuracy of the HPFs elicited under various protocols.

A. The standard hypothetical choice protocol

Consistent with the literature on stated preferences, Table 2 documents substantial hypothetical bias: the standard-protocol HPF overstates the RPF by nearly 7 percentage points overall (equivalently, by 28.6%), and we reject the absence of bias ($p < 0.001$). Nevertheless, there is a strong correlation across items between the RPF and the HPF ($\rho = 0.697$), which suggests that this HPF may be a useful predictor of the RPF, even if it is not a good prediction.

Figure 1 shows a scatterplot of the RPF against the standard-protocol HPF. We have pooled all data, so there are two observations for each item (one for \$0.25 and one for \$0.75).

Due to the overall bias noted above, most of the data points (73%) lie below the 45-degree line. However, the strong correlation between the RPF and the HPF is also readily evident.

As shown in Table 2, the variance of the HPF is more than twice that of the RPF. Thus, we find evidence not only of hypothetical bias, but also of what we will call *hypothetical noise*.[17] It is natural to 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. Performing the same decomposition as for the RPFs, we have:

$$E\left[s_{Hj}^2\right] = \sigma_{Hj}^2 + \sigma_{\mu j}^2(1-\rho_{Hj}),$$

where all the terms are analogous to those defined for the RPFs, and μ (rather than ω) denotes the sampling error. Greater "randomness" in choice can show up as HPFs that are closer to 50 percent than the RPFs (which increases $\sigma_{\mu j}^2$), and/or less correlation between the sampling error for distinct item-price pairs ($\rho_{Hj} < \rho_{Rj}$). Still, by the same reasoning as for the RPFs, $\bar{R}_j^P(1-\bar{R}_j^P)/N$ provides an upper bound on $\sigma_{\mu j}^2(1-\rho_{Hj})$. Thus, $\sigma_{\mu,0.25}^2(1-\rho_{Hj}) < \frac{38.33(100-38.33)}{28} = 84.4$, and $\sigma_{\mu,0.75}^2(1-\rho_{Hj}) < \frac{23.44(100-23.44)}{28} = 64.1$. But then, because $s_{H,0.25}^2 = 222.8$, we infer that $\sigma_{H,0.25}^2 > 138.4$; likewise, because $s_{H,0.75}^2 = 158.8$, we infer that $\sigma_{H,0.75}^2 > 94.7$. Those lower bounds exceed, respectively, $s_{R,0.25}^2 = 120.7$ and $s_{R,0.75}^2 = 83.2$. Because $\sigma_{R,0.25}^2$ and $\sigma_{R,0.75}^2$ are likely considerably smaller than the latter figures (which include sampling error), we conclude that σ_{Hj}^2 likely exceeds σ_{Rj}^2 by a wide margin.

It follows from this calculation that the phenomenon of 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, Table 2 shows that this is the case with price variation. This result is encouraging: it suggests that 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.

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. The right half of Table 2 contains performance statistics for the difference

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

method, which generally underperforms the myopic benchmark. The overall bias is roughly the same in absolute value: differencing the HPF overstates the actual price response by a factor of just under two. The MSPE is substantially larger than for the myopic benchmark, and calibration noticeably poorer. The calibration parameter for changes, for which the myopic benchmark provides no counterpart, is extremely low (0.248).

Table 3 contains performance statistics for the levels method. For the most part, performance remains poor relative to the myopic benchmark. The one exception is a lower overall bias (3.22 percentage points) when predicting from \$0.25 to \$0.75 choices; but this bias is still nearly half of the actual price change. The approach does noticeably better with respect to overall bias and MSPE predicting from \$0.25 to \$0.75 choices than the other way around, but not with respect to calibration. Once again, the calibration parameters for changes, for which the myopic benchmark provides no counterpart, are quite low (0.207 and 0.240).

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 present 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 double the size of the sample used to compute the HPFs by combining the H and HD treatments, thereby significantly reducing sampling error. For the second, we estimate new versions of equation (1), using HD-treatment versions of right-hand-side variable as instruments for the H-treatment versions. Because the H-treatment and HD-treatment versions of the HPF for

[18] Sampling error in the measurement of the RPF should not matter with calibration for levels using the levels approach or with 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 matter with calibration for changes using the levels approach and with 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.

any item reflect the same population tendencies, they are necessarily correlated, and because they reflect independent random draws from the population, their sampling errors are 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.

B. Alternative hypothetical choice protocols

As we mentioned in Section 2, studies in the SP literature gather hypothetical choice data using a variety of protocols, some of which are intended to "fix" the standard hypothetical choice question. Here we examine the accuracy of hypothetical purchase frequencies elicited through other protocols. Two of those protocols elicit purchase likelihoods, either for the respondent (treatment HL) or for a typical undergraduate of the same gender (treatment HV). In each case, we create two HPF measures, classifying a response as a purchase if it indicates, respectively, certainty (i.e., a "1") or high likelihood (i.e., either a "1" or a "2"). We label these alternatives "likely (1)," "3rd party (1)," "likely (≤2)," and "3rd party (≤2)." We also create two alternative HPFs using the hypothetical WTPs. For one, we treat a response as indicating a purchase if the WTP exceeds the price. For the other, we follow the spirit of the procedure NOAA considered: we multiply the stated WTP by an adjustment factor, and then compare the adjusted WTP to price. We choose the adjustment factor so that the implied HPF coincides as close as possible with the RPF for the calibration sample.[20] For the summary statistics reported in Table 2, we use all of our data to calibrate the adjustment factor. However, when predicting from choices at price P_1 to choices at price P_2, we use only the choices at price P_1 as the calibration sample (because the exercise assume that data on real choices at the price P_2 are unavailable). Results appear in Tables 2 and 3. Based on these results, we reach the following conclusions.

First, consistent with findings in the literature, several alternative protocols reduce the overall degree of hypothetical bias (shown in the second column of Table 2). Ignoring the adjusted WTP (for which low bias is guaranteed by construction), the cheap-talk protocol performs best according to this metric, followed closely by "3rd party (1)." For the cheap-talk

[20] Because the distribution of WTPs is "lumpy," it is usually impossible to find an adjustment for which the actual and implied purchase frequencies match exactly.

protocol, the gap between the average RPF and the average HPF falls from 6.86 to 2.45 percentage points (though it remains statistically significant, $p < 0.01$).

There are, however, two possible explanations for results such as these: one is that the protocol mitigates the cause of the bias; the other is that it introduces an offsetting bias. If the second explanation is correct, then the putative benefit of the protocol may reflect a fortunate coincidence rather than a legitimate solution. Significantly, that explanation would account for the observation that the performance of the cheap-talk protocol has proven somewhat sensitive to its details and to the context. Additional results described below provide several reasons to credit the second explanation rather than the first, and hence to question the value of alternative protocols that appear to reduce hypothetical bias.

Second, of the approaches we consider, the one that arguably performs best overall is "3rd party (1)."[21] By almost all of the metrics, it is either the best or one of the best alternatives. It performs especially well when used in conjunction with the difference method (Table 2): in that case, the bias in predicting the change in the RPF is 1.65 percentage points, less than one-quarter of the bias for the standard protocol; the MSPE, 68.2, is roughly half as large for the standard method; and the calibration coefficients for levels, 0.681 (predicting \$0.25 choices) and 0.566 (predicting \$0.75 choices), are higher than for all other alternatives. Only the calibration coefficient for changes (0.217) is inferior to those associated with some of the other methods. The superior performance of this approach does not surprise us, in that questions about third parties do not trigger motives pertaining to social image that can create divergences between responses to hypothetical and real choice questions. These findings are notable in that, to our knowledge, vicarious hypothetical choice questions have not been used in the SP literature.

While the "3rd party (1)" approach performs well relative to other alternatives involving hypothetical choices, its performance may not be "good enough" for economic applications. For example, with the difference method, the average error of 1.65 percentage points represents more than 20% of the actual change in the RPF. Accuracy is even lower with the levels method: when predicting from \$0.25 to \$0.75 choices, the average error is -3.33, more than 40% of the actual change. We will return to this important issue below.

Third, aside from the "3rd party (1)" approach, none of the alternatives considered yields a clear improvement over the standard hypothetical choice protocol. Strikingly, the overall correlation between the RPF and the standard-protocol HPF is higher than for any alternative

[21] We say "arguably" because the comparison hinges on how one weights the various performance metrics.

HPF, which casts doubt on the hypothesis that the alternative protocols improve the informational content of the hypothetical choice measures.

In other respects, comparisons between the standard protocol and the other alternatives (aside from "3rd party (1)") are decidedly mixed. Take the cheap-talk protocol. The summary statistics in Table 2 show that, despite achieving the lowest overall hypothetical bias, it slightly reduces the correlation between the RPF and the HPF, and slightly increases variance (an indicator of hypothetical noise). When using the difference method to make predictions (also Table 2), the cheap-talk protocol *amplifies* hypothetical bias: on average, the predicted price response is roughly two-and-a-half times as large as the actual response (versus two times for the standard protocol). It also noticeably underperforms the standard protocol with respect to MSPE, but performs modestly better with respect to calibration. Using the levels method, the two approaches are almost identical with respect to all metrics when predicting choices at $0.75, but the cheap-talk approach performs somewhat better when predicting choices at $0.25.

The story is similar for the other alternatives. In most cases, whether a given alternative is better or worse than the standard protocol depends on the method used to make predictions (differences or levels), and how one weights the various performance metrics. Aside from "3rd party (1)," no other approach yields a clear improvement. Accordingly, these results suggest that the other protocols reduce hypothetical bias mainly by inducing offsetting biases, rather than by curing the causes of the bias. Whether they improve or degrade the informational content of hypothetical choices remains somewhat unclear; we return to that question in Section 7.

Fourth, all of the alternative protocols perform poorly relative to appropriate benchmarks. First consider comparisons to the myopic benchmark; for convenience, we reproduce its performance statistics in Tables 2 and 3. We reiterate that this is a very low standard – any method that underperforms myopia merits no further consideration. The various alternative HPFs outperform this benchmark in some cases, depending on the method used (differences or levels) and whether one is forecasting choices at $0.25 or $0.75. However, the only method that consistently improves upon it (regardless of method or direction of the forecast) is "3rd party (1)." With respect to MSPE, improvements are uncommon (only three of the 25 cases shown in the tables), and in most cases performance deteriorates considerably. The "3rd party (1)" approach accounts for two of the three instances of improvement: it performs quite a bit better when using the differences method, a bit better when using the levels method to predict choices at $0.75, and a bit worse when using the levels method to predict choices at $0.25 (averaging about the same across the two directions). The only other improvement with respect to MSPE is for "likely (2)"

which, given the other performance statistics, appears unrepresentative. With respect to calibration for levels, all of the methods significantly underperform the myopic benchmark. The only exception involves the use of adjusted WTP to predict choices at $0.75. That result is plainly an outlier for the adjusted WTP method, which generally falls far short of the myopic benchmark. Calibration for differences is not defined for this benchmark.

Thus, the only alternative that arguably yields a significant and consistent improvement over the myopic benchmark couples the "3rd party (1)" approach with the difference method. That alternative achieves significant reductions in bias and MSPE at the cost of somewhat poorer calibration. However, as noted above, even that alternative yields an average prediction error exceeding 20% for price sensitivity. Furthermore, its performance falls far short of the most demanding benchmark shown in Table 1, which provides an indication of what more standard methods achieve when better choice data are available. Consequently, even the best alternative considered in Tables 2 and 3 may not merit serious consideration as a tool for predicting behavioral responses to changes in economic parameters.

Fifth, good performance with respect to calibration for differences is particularly hard to achieve. Using the difference method, the largest calibration parameter for differences is 0.272; for the levels method, it is 0.251 when predicting $0.25 choices, and 0.316 when predicting $0.75 choices. Thus, the variation in actual price sensitivity across items is only weakly related to the variation in predicted price sensitivity, regardless of the protocol and method used. Sampling error in the HPFs is part of the explanation, but as we saw at the end of the last subsection, increases in group size improve calibration only to a limited degree.

Sixth, predictions based on hypothetical WTPs are particularly poor. Though a NOAA-style adjustment minimizes hypothetical bias within sample by construction, it is of practically no value when predicting out of sample using the difference method. Though it improves performance somewhat when predicting out of sample using the levels method, the average bias and MSPE remain sizeable. Thus, whether or not one makes a NOAA-style adjustment, the use of stated WTPs leads to remarkably poor out-of-sample predictions.

7. The accuracy of predictions based on hypothetical choices and non-choice ratings

Next we evaluate the accuracy of predictions based on statistical relationships between actual choices, hypothetical choices, and non-choice ratings. In contrast to the preceding section, here we treat HPFs as predictors rather than predictions. Specifically, we estimate models relating RPFs to HPFs and other non-choice ratings using data for a single price, P_1 (either $0.25

or \$0.75). Then we use those models along with additional data on hypothetical choices and non-choice ratings to predict RPFs at the alternative price, P_2. As in the previous section, one can construct these predictions using either the levels method or the difference method. For the levels method, we simply set $\hat{R}_{i2} = R^F_{i2}$ (where R^F_{i2} denotes the fitted value of $\hat{R}_{i2}$ based on the model); for the difference method, we set $\hat{R}_{i2} = R_{i2} + \left(R^F_{i2} - R^F_{i1}\right)$. For the sake of brevity, we report results based only on the difference method. In conducting our analysis, we found that, for model-based predictions, the difference method almost always outperformed the levels method.

A key step in building good predictive models is model selection. 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-of-sample predictions. We examine several selection criteria. One is the AIC (Akaike Information Criterion), a measure of goodness-of-fit that includes a penalty based on the number of parameters in the model,[22] commonly used for model selection when accurate out-of-sample prediction is the objective.[23] As in Section 4, we also use the LASSO procedure, which performs model selection automatically by maximizing the penalized sum of squared residuals.[24] Additional selection criteria involve cross-validated measures of predictive performance, such as MSPE.

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 (treatment HD) is above or below the median.[25] To understand why, recall that our out-of-sample predictions either 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.

A. Predictions based on simple models

We begin by examining the predictive performances of simple univariate OLS regressions of the RPF on the various HPF, one at a time. Because the RPF aggregates binary

[22] 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.
[23] Results based on another well-known alternative, the BIC (Bayesian Information Criterion) are similar.
[24] The penalty is proportional to the size of the coefficient vector in the L_1 norm; its weight is determined by minimizing cross-validated MSPE.
[25] 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.

choices over subjects, there are potential justifications for employing other specifications. However, a look at the scatterplot shown in Figure 1 suggests that a linear function will likely fit the data well. That impression is confirmed by some non-parametric estimates shown in the Appendix. It is also important to bear in mind that our objective here is to estimate predictive relationships rather than causal relationships. As White (1980) has shown, predictions based on OLS estimates always yield the lowest expected MSPE conditional on using the adopted specification, even when that specification deviates from the true functional form.

Table 4 reports model selection criteria for these univariate models. When predicting from \$0.75 choices, both the AIC and the CV-MSPE favor using a model that incorporates the standard HPF over all other alternatives. When predicting from \$0.25 choices, the AIC favors the cheap-talk HPF followed by the standard HPF, while the CV-MSPE favors the standard HPF followed by the cheap-talk HPF.

Table 5 reports our various metrics of out-of-sample predictive accuracy for the univariate models. At the top of the table, we also reproduce some benchmark results.

The first lesson from Table 5 is that a simple regression of the RPF on the standard HPF yields an equation that performs admirably with respect to predicting the purchase frequencies that would be observed after a large price change. The average biases are quite small: −0.55 when predicting from \$0.75 choices, and 0.44 when predicting from \$0.25 choices. These errors represent only 7.3% of the actual average price response in the first instance and 5.9% in the second – in each case, well within the tolerances to which economists are accustomed. In terms of MSPE, this specification outperforms the myopic benchmark by a wide margin; more impressively, it matches the more challenging benchmarks (which use data on real choices at both prices) when predicting from \$0.75 choices, and is at least in the same ballpark when predicting from \$0.25 choices. Calibration is much improved compared with the results in Tables 2 and 3. For levels, the calibration parameter is 0.919 when predicting from \$0.75 choices, which falls a bit short of the benchmarks but nevertheless is nearly ideal; when predicting from \$0.25 choices, it is 0.692, which at least surpasses the myopic benchmark. Given the difficulty of achieving good calibration for changes (see the results for the RPE predictor benchmark), the associated parameters (0.531 and 0.523) are respectable, although clearly there is room for improvement.

This simple regression equation 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 the Chow test statistic reported in Table 5 (p = 0.593), 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. 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 choices and the \$0.75 choices on the figure, along with error bands. For all practical purposes, they are indistinguishable. 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 the Appendix).

The second lesson to be drawn from Table 5 is that, when HPFs are used as predictors rather than predictions, the standard protocol is generally superior to the alternatives. We find this result surprising in light of the literature on methods for improving hypothetical questions, though obviously less so given Tables 2, 3, and 4. Specifications using the cheap-talk HPF yield some improvement in the MSPE when predicting from \$0.75 choices, as well as in two of the calibration parameters. However, these gains come at the cost of substantially greater bias, which reflects the fact that the Chow test rejects equality of the coefficients across the \$0.25 and \$0.75 samples (p = 0.042). Specifications using the "likely (1)" variable achieve very low bias; not surprisingly, the Chow test statistic fails to reject equality of the coefficients (p = 0.724). However, the MSPEs are significantly higher and the calibration parameters lower. Specifications using the "likely (2)" variable achieve a small reduction in bias when predicting from \$0.25 choices, but no other gains. Surprisingly, specifications using the "3rd party (1)" variable yield no improvements, and those using the "3rd party ($\leq$2)" variable only improve one of the calibration parameters. Finally, when predicting from \$0.75 choices, the specification that uses the WTP variable improves one of the calibration parameters and generates predictions with virtually no bias. However, when predicting in the opposite direction, the bias is quite large, MSPE rises, and the other calibration parameters decline.

The third lesson to take from Table 5 is that the univariate prediction approach works tolerably well for *all* of the protocols. Relative to the myopic benchmark, the average bias falls by more than 50% in all cases but one (for which it also declines), MSPE falls by more than 40% in all cases but one (for which it rises), and calibration in levels is generally comparable (though a bit lower when predicting from \$0.75 choices). Though in many instances predictive performance falls short of the more demanding benchmarks (that make use of additional choice data), it is generally closer to those standards than to myopia.

Next we ask whether it is possible to improve out-of-sample predictive accuracy by using more than one HPF and other non-choice ratings in combination. We would expect specifications that include multiple HPFs to yield more accurate predictions if the alternative protocols elicit different types of predictively useful information (as opposed to measuring the same information with different noise). In addition, if (as intended) the questions posed to subjects participating in treatment S address the likely causes of divergences between RPFs and HPFs, we would expect to achieve further improvements by including measures of the associated responses.

To determine whether the treatment S data capture some of the causes of hypothetical bias, we estimate a collection of bivariate regressions, each of which relates the RPF to the standard HPF and the mean response for one of the treatment S questions, pooling all of our data.[26] Regression results appear in Table A.1 of the Appendix. The coefficients of the additional non-choice rating variables are all highly statistically significant, with the exception of the temptation variable. Accordingly, it appears likely that, by exploiting the information contained in the additional rating variables, we should be able to improve upon predictions that use only hypothetical choice variables.

Rather than consider all possible permutations of predictors, for the remainder of this section we will include the standard HPF in all specifications (on the grounds that it is arguably the best single predictor), and examine the effect of adding each of the other HPFs and non-choice rating variables, one at a time. Table 6 reports model selection criteria for these bivariate models (and reproduces corresponding statistics for the best univariate specification).

The inclusion of a second HPF improves the AIC in all cases, and it improves the CV-MSPE in all but one. The preferred co-predictor among the HPFs when predicting from the \$0.75 choices is the "3rd party (1)" HPF according to both the AIC and the CV-MSPE; when predicting from the \$0.25 choices, it is the cheap talk HPF according to the AIC, and the "3rd party (1)" HPF according to the CV-MSPE. The inclusion of rating variables yields improvements in some but not all cases. The preferred co-predictor among the ratings is the liking variable according to both criteria, regardless of whether one predicts \$0.25 or \$0.75 choices. Overall, the preferred

[26] A seemingly natural alternative would have been to regress the difference between the RPF and the HPF on the same variables. However, we know from Figure 1 that the magnitude of hypothetical bias increases (both absolutely and proportionately) with the purchase frequency. As a result, any variable that is correlated with the desirability of the item will appear to account for the gap. Moving the HPF to the right-hand side of the equation is more appropriate for our purposes, because our object is to determine whether the non-choice ratings can be used to improve the *best* prediction one can make based on the HPF.
Yet another alternative would have been to estimate a single regression with the HPF and all the non-choice ratings on the right-hand side. That strategy would certainly be more appropriate were we primarily interested in causal interpretations of the coefficients. However, our main objective is to assess incremental contributions to predictive power. While we will eventually search for the best combination of predictors, it is useful to start by examining their performances one at a time.

co-predictor is the "3rd party (1)" HPF according to both criteria when predicting from $0.25 choices, and the liking variable according to both criteria when predicting from $0.75 choices.

There is, however, an important caveat with respect to the apparent implication of Table 6 that one can improve predictive performance by using the standard HPF in combination with other hypothetical choice and non-choice rating variables. As we have noted, our HPFs are measured with sampling error. Thus, we would not be surprised to see improvements like those in Table 6 even if the additional variables were nothing more than noisy proxies for standard-protocol HPFs elicited with new groups of subjects. If one has the opportunity to gather data from additional subjects, it is therefore unclear whether one should enlarge the standard-protocol sample, or collect different types of non-choice data from a new sample.

To shed light on this issue, we also evaluate a bivariate specification containing the HPFs for treatments H and HD (both of which use the standard protocol). Results appear in last line of Table 6 (labeled "Hypothetical – duplicate"). Notice that this specification is preferred to all others according to both criteria when predicting from $0.75 choices, and according to the CV-MSPE when predicting from $0.25 choices (in which case it is also a close second according to the AIC). Thus, our within-sample criteria favor enlarging treatment group H over the alternatives. One should bear in mind, however, that the benefits of gathering more data using the same protocol decline with the size of the treatment group, because the sampling error shrinks. Thus, with larger treatment samples, the benefits of adding hypothetical and non-choice rating variables would likely be even more apparent according to our within-sample criteria.

Table 7 reports our various metrics of out-of-sample predictive accuracy for the bivariate models that include two HPFs. For convenience, at the top of the table, we also reproduce results for some key benchmarks and the for the preferred univariate model (which includes only the standard HPF). The model that includes the standard HPF plus the "3rd party (1)" HPF, which our within-sample model selection criteria generally favor over specifications that add other alternative HPFs, outperforms the preferred univariate model across the board. The average biases are tiny: 0.16 when predicting from $0.75 choices, and −0.19 when predicting from $0.25 choices; not surprisingly, the Chow test statistic fails to reject equality of the coefficients, which are virtually the same for the two subsamples (p = 0.937). These biases represent only 2.1% of the actual average price response in the first instance and 2.5% in the second – acceptable margins of error even by the most exacting standards. This specification also achieves a lower MSPE than even the most demanding benchmarks (those that use additional choice data) when predicting from $0.75 choices, and underperforms them by only a slightly larger margin when

predicting from $0.25 choices. Finally, the calibration parameters are all respectable (0.924 and 0.728 for levels, and 0.645 and 0.674 for differences), if still somewhat lower than for the best-performing benchmark. Thus, this simple specification yields remarkably accurate predictions of the purchase frequencies that would be observed after a large price change.

Other specifications in Table 7 yield predictions that improve upon those obtained from the preferred univariate regression by one or more criteria, particularly MSPE and calibration. However, these gains often come at the cost of greater bias. Significantly, with respect to bias and MSPE, *all* of these specifications – including the one that adds a second standard HPF based on treatment HD – uniformly underperform the one that includes the standard HPF and the "3rd party (1)" HPF; moreover, none performs much better with respect to any aspect of calibration.

Table 8, which is configured identically to Table 7, reports our metrics of out-of-sample predictive accuracy for bivariate models that include the standard HPF along with one of our non-choice ratings variables. The model that includes both the standard HPF and the liking variable, which our within-sample model selection criteria favor within this group, delivers excellent calibration parameters (0.987 and 0.722 for levels, and 0.783 and 1.064 for differences), but also produces substantial absolute biases, which drive up MSPE. As indicated by the Chow test statistic, the coefficients of the estimated relationship differ significantly between the two subsamples, and in this instance those differences are consequential.

Significantly, every bivariate model in Table 8 improves every measure of calibration relative to the preferred univariate model. Thus, including non-choice ratings in the set of predictors may be the key to achieving high-quality calibration. MSPE also falls for a number of the specifications, with the largest declines occurring for the ones that add the "approve/disapprove" and "happiness" variables; indeed, both of those specifications arguably perform better overall than the one that includes a second standard HPF based on treatment HD. Notably, among the specifications that add a non-choice ratings variable, those rank second and third according to our within-sample model selection criteria in three of four cases, and one ranks second in the fourth case (see Table 6). Many of the specifications in Table 8 yield larger biases than the preferred univariate specification, though the difference is modest in several cases.

B. Predictions based on optimized models

There is, of course, no reason to restrict our models to one or two predictors. However, with no restrictions on the number of predictors, the set of possible models becomes enormous. Here we exacerbate that problem by greatly expanding the set of predictors. Specifically, for questions that can elicit more than two distinct responses, we construct variables measuring the

frequency of each response (leaving one out because the frequencies sum to one). We also include squares of the HPFs and average ratings, as well as a collection of interactions. Thus, we draw on an extremely large set of potential predictors.

As mentioned in Section 3, 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 hypothetical choices and non-choice ratings.

Table 9 reports our metrics of out-of-sample predictive accuracy for the resulting LASSO model. At the top of the table, we also reproduce results for some key benchmarks, as well as for the preferred bivariate model (i.e., the one that includes both the standard and "3rd party (1)" HPFs). The LASSO model improves upon the preferred bivariate model with respect to MSPE and calibration. Some of those improvements are substantial. The largest improvements pertain to the calibration parameters for differences, which we previously noted were the weakest aspects of the bivariate model's performance. The absolute value of the prediction error is roughly unchanged when predicting from \$0.75 choices (and represents an average error equal to only 2.7% of the actual demand response); however, it increases from 0.19 to 0.69 (or 9.2% of the actual demand response) when predicting from \$0.25 choices. Overall, this model's performance compares well with that of even the most demanding benchmark (which employs additional real choice data), surpassing its performance in several dimensions.

We can potentially achieve further improvements by fine-tuning our model selection criteria. Specifically, using cross-validation as in Tables 4 and 6, we can identify OLS models that perform well when predicting from choices involving relatively attractive options to choices involving relatively unattractive ones, and the other way around. We accomplish that objective by employing a hill-climbing algorithm to find the OLS model that optimizes a specified measure 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. Because we evaluate out-of-sample predictive performance according to three distinct criteria (MSPE, average bias, and calibration), we conduct one search to find the model that minimizes cross-validated MSPE, another to minimize absolute mean prediction error (AMPE), and a third to minimize the distance between the cross-validated calibration parameter and unity.

Metrics of out-of-sample predictive accuracy for the resulting models also appear in Table 9. Minimizing cross-validated MSPE or AMPE generally degrades predictive

performance. However, maximizing cross-validated calibration quality yields models that surpass even the LASSO specifications. The associated average biases are trivial: 0.10 (or 1.3% of the demand response) when predicting from $0.75 choices, and −0.16 (or 2.1% of the demand response) when predicting from $0.25 choices. The resulting models also achieve meaningful reductions in MSPE, as well as a significant gain in calibration for differences when predicting from $0.75 choices; one of the other calibration parameters increases slightly and the other two decrease slightly. Depending on how one weights the various metrics, it is arguable that this specification meaningfully outperforms even the most demanding benchmark.

It is natural to wonder whether it is necessary to consider such a large class of potential predictors to achieve such high standards of predictive performance. As alternatives, we estimate a LASSO specification that only draws on all the hypothetical choice variables, and one that only draws on the standard-protocol HPF along with the all the non-choice ratings variables. (A specification that draws only on the ratings variables and none of the hypothetical choice variables performs so poorly that it is not worth considering.) Results appear in the bottom portion of Table 9. In both instances, the restriction on the set of potential predictors significantly degrades predictive performance, mostly by introducing substantial biases. Notice that the LASSO procedure does *not* select the variables in our preferred binary model when it is confined to hypothetical choice variables, despite the fact that the out-of-sample predictive performance of that model exceeds that of the model that LASSO does select. That finding underscores the difficulty of identifying within-sample model selection criteria that assure good out-of-sample predictive performance.

So far, we have explored the predictive power of specifications that use only variables derived from hypothetical and ratings questions. We now expand our analysis to include variables that measure the physical characteristics of the items.[27] Our purpose is to determine whether one can achieve further improvements in predictive accuracy by adding physical characteristics to the list of potential predictors. It is important to emphasize that any such improvements potentially come at significant 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 context-specific predictive models. For that purpose, it is important to use predictors for which meaning does not vary over the

[27] 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.

intended domain. One extremely important advantage of an HPF is that its meaning is universal across all domains. For the most part, we have also focused on non-choice reactions for which meaning is 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, 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 a LASSO specification that draws on all variables measuring hypothetical choices, non-choice ratings, and physical characteristics appear in the second-to-last line of Table 9. The calibration parameters are superb: 0.972 and 0.771 for levels, and 0.968 and 0.902 for differences. Relative to the LASSO models that exclude all physical characteristics, MSPE rises when predicting from $0.75 choices and declines when predicting from $0.25 choices. Unfortunately, bias increases to a troubling degree when predicting from $0.75 choices.

These findings admit at least two interpretations. One is that any sacrifice in predictive accuracy resulting from eschewing context-specific variables in order to enhance the model's potential portability across domains is relatively small. The other is that a context-specific prediction model can achieve nearly ideal calibration; conceivably, through the use of more refined model selection criteria, these improvements in calibration could be achieved without increasing average bias.

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 instead no price variation is observed but non-choice response variables are available. We close 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 10 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 third 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. Thus, the use of non-choice response variables significantly enhances predictive performance even when good choice data are available.

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 more conventional approaches. While hypothetical choice frequencies are poor predictions of real choice frequencies irrespective of the elicitation protocol, they are nevertheless excellent predictors, particularly when used in combination with each other 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 are especially thorny. Further exploration of within-sample selection criteria is warranted, and we are far from exhausting the range of questions that might yield valuable predictors. Other largely 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. Also, if the methods explored in this paper are to be of any practical value, it will be necessary to resolve a variety of pragmatic and conceptual issues concerning their use in real applications, as opposed to the laboratory.

There are, of course, real-world contexts in which nominally hypothetical questions are either consequential or perceived as such, and where 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, an important potential advantage of this strategy over conventional methods of predicting choices in as-yet unobserved situations is that it requires a model of only a single process (one determining how non-choice reactions are related to real choices), rather than a distinct model for every decision context.

References

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.

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.

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.

Gruber, Jonathan, and Ebonya Washington, "Subsidies to employee health insurance premiums and the health insurance market," *Journal of Health Economics* 24 (2), 2005, 253-276.

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,” mimeo, Princeton University, 2011.

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

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," mimeo, 2012.

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.

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.

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

Table 1: Predictive accuracy for benchmark methods

Benchmark method	**Predicting from $0.75 to $0.25**				**Predicting from $0.25 to $0.75**			
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)
Methods using limited choice data								
Myopic	−7.50	93.3	1.001	NA	7.50	93.3	0.690	NA
Structural: Z_i = g/serv	13.61	298.5	0.713	-0.004	-16.12	339.7	4.432	0.318
Structural: Z_i = g/serv, nut/serv	15.04	335.1	0.708	0.002	-16.12	338.1	3.868	0.322
Structural: Z_i = g/serv, cat*g/serv	11.57	239.6	0.844	0.033	-14.31	281.7	2.638	0.322
Structural: Z_i = g/serv, nut/serv, cat*g/serv	15.69	341.2	0.808	0.069	-16.58	350.4	4.189	0.326
Methods using additional choice data								
Average Δ	0	37.2	1.000	NA	0	37.2	0.690	NA
RPF predictor	0	37.3	0.994	-3.388	0	26.4	1.000	0.946
Augmented predictors: OLS	0.19	40.8	0.950	0.265	-0.26	37.9	0.795	0.486
Augmented predictors: Lasso	0.04	35.9	1.021	0.780	-0.02	26.4	1.053	0.835

Table 2: Measures of hypothetical demand: summary statistics and predictive accuracy of the difference method

Demand Variable	Summary statistics				Predictive accuracy, difference method				
	Mean (%)	Overall hyp. bias	Variance	Correlation with RPF	Average bias	MSPE	Calibration (level, to $0.25)	Calibration (level, to $0.75)	Calibration (Δ)
Real	24.01	0	115.7	1	NA	NA	NA	NA	NA
Myopic benchmark	NA	NA	NA	NA	-7.50	93.3	1.001	0.690	NA
Hypothetical	30.88	6.86	245.8	0.697	7.39	137.4	0.610	0.500	0.248
Hyp - cheap talk	26.46	2.45	254.0	0.693	11.34	206.2	0.612	0.518	0.272
Hyp – likely (1)	17.89	-6.13	147.5	0.635	5.37	108.9	0.605	0.512	0.140
Hyp – likely (≤2)	29.97	5.96	276.7	0.666	11.10	218.0	0.566	0.463	0.194
Hyp – 3rd party (1)	21.50	-2.51	145.4	0.643	1.65	68.2	0.681	0.566	0.217
Hyp – 3rd party (≤2)	43.47	19.46	264.1	0.582	4.58	119.1	0.582	0.460	0.010
Hyp – WTP	64.20	40.18	358.7	0.594	18.41	494.1	0.402	0.369	0.062
Adjusted hyp – WTP	23.27	-0.74	495.5	0.511					
From $0.75 to $0.25					40.48	1776.2	0.465		0.135
From $0.25 to $0.75					14.13	406.2		0.152	0.023

Table 3: Measures of hypothetical demand: predictive accuracy of the levels method

Demand Variable	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75			
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)
Myopic benchmark	−7.50	93.3	1.001	NA	−7.50	93.3	0.690	NA
Hypothetical	10.56	243.1	0.474	0.207	3.17	103.4	0.466	0.240
Hyp - cheap talk	8.12	171.3	0.538	0.251	-3.22	104.5	0.458	0.242
Hyp – likely (1)	-3.44	120.5	0.538	0.169	-8.81	148.7	0.593	0.308
Hyp – likely (≤2)	11.51	292.2	0.419	0.171	0.41	85.9	0.489	0.246
Hyp – 3rd party (1)	-1.68	115.3	0.525	0.148	-3.33	85.6	0.542	0.316
Ed on Hyp – 3rd party (≤2)	21.75	670.0	0.345	0.095	17.17	438.9	0.351	0.131
Hyp – WTP	49.39	2546	0.551	0.177	30.98	1147.2	0.29	0.13
Adjusted hyp – WTP	38.44	1620.3	0.445	0.151	-18.81	427.9	0.943	0.33

Table 4: Model selection, specifications employing a single hypothetical choice variable

Hypothetical Choice Variable	**Predicting from $0.75 to $0.25**		**Predicting from $0.25 to $0.75**	
	AIC	CV-MSPE	AIC	CV-MSPE
Hypothetical	-465.2	51.0	-394.8	64.8
Hyp - cheap talk	-442.3	64.2	-413.3	69.7
Hyp – likely (1)	-422.6	78.0	-377.6	79.7
Hyp – likely (≤2)	-454.9	56.6	-378.2	77.5
Hyp – 3rd party (1)	-445.3	54.7	-373.1	74.7
Hyp – 3rd party (≤2)	-433.2	53.5	-346.7	79.9
Hyp – WTP	-418.9	76.2	-369.3	84.1

Table 5: Predictive accuracy of specifications employing a single hypothetical choice variable

Model	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75				Chow Test
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)	
Benchmarks									
Myopia	−7.50	93.3	1.001	NA	−7.50	93.3	0.690	NA	NA
RPF predictor	0	37.3	0.994	-3.388	0	26.4	1.000	0.946	NA
Augmented predictors: LASSO	0.04	35.9	1.021	0.750	−0.02	26.4	1.053	0.835	NA
Hypothetical	-0.55	36.2	0.919	0.531	0.44	36.4	0.692	0.523	0.593
Hyp - cheap talk	1.13	34.8	0.895	0.595	-2.63	43.7	0.708	0.507	0.042
Hyp – likely (1)	0.14	48.2	0.771	0.236	0.57	45.7	0.683	0.260	0.724
Hyp – likely (≤2)	1.59	44.3	0.850	0.398	-0.30	38.4	0.692	0.463	0.451
Hyp – 3rd party (1)	-2.54	46.5	0.873	0.401	2.69	46.8	0.692	0.413	0.004
Hyp – 3rd party (≤2)	-3.25	54.9	0.938	0.029	3.33	55.2	0.648	0.030	0.001
Hyp – WTP	0.01	43.6	0.961	0.214	-6.78	114.7	0.540	0.112	0.056

Table 6: Model selection, specifications employing two predictors

Predictors	Predicting from $0.75 to $0.25		Predicting from $0.25 to $0.75	
	AIC	CV-MSPE	AIC	CV-MSPE
Hypothetical	-465.2	51.0	-394.8	64.8
Hypothetical plus:				
Hyp - cheap talk	−470.3	50.0	−424.8	59.6
Hyp – likely (1)	−472.1	53.5	−409.6	58.8
Hyp – likely (≤2)	−479.2	46.8	−406.9	60.7
Hyp – 3rd party (1)	−488.3	40.6	−407.7	57.0
Hyp – 3rd party (≤2)	−477.0	42.6	−396.5	61.5
Hyp – WTP	−469.4	50.4	−406.2	59.7
Approve/disapprove	−473.6	52.4	−421.3	57.7
Happiness	−466.3	49.6	−408.4	57.2
Liking	−476.2	47.1	−426.2	56.8
Regret	−463.2	52.2	−405.6	64.9
Tempting	−463.8	50.8	−396.1	64.0
Enjoy if harmless	−464.4	50.6	−397.9	64.2
Good/bad for you	−463.3	53.2	−403.9	69.2
Pos/neg impression	−463.2	52.1	−403.4	66.7
Over/understate	−464.9	54.4	−405.6	70.9
Hyp – duplicate	−495.1	38.8	−424.5	55.8

Table 7: Predictive accuracy of specifications employing two types of hypothetical choice frequencies

Model	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75				Chow Test
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)	
Benchmarks									
Myopia	−7.50	93.3	1.001	NA	−7.50	93.3	0.690	NA	NA
RPF predictor	0	37.3	0.994	-3.388	0	26.4	1.000	0.946	NA
Augmented predictors: LASSO	0.04	35.9	1.021	0.750	-0.02	26.4	1.053	0.835	NA
Hyp - preferred univariate	−0.55	36.2	0.919	0.531	−0.44	36.4	0.692	0.523	0.860
Hypothetical plus:									
Hyp - cheap talk	0.94	32.5	0.930	0.650	2.56	38.8	0.730	0.609	0.041
Hyp – likely (1)	1.15	36.7	0.879	0.545	0.82	35.5	0.726	0.567	0.644
Hyp – likely (≤2)	1.63	35.9	0.919	0.604	1.08	33.7	0.720	0.631	0.531
Hyp – 3rd party (1)	0.16	32.3	0.924	0.645	−0.19	31.8	0.728	0.674	0.937
Hyp – 3rd party (≤2)	−0.28	34.6	0.961	0.596	−0.25	34.8	0.699	0.575	0.784
Hyp – WTP	0.91	35.5	0.955	0.574	4.24	58.5	0.657	0.427	0.049
Hyp – duplicate	0.70	33.7	0.936	0.600	0.90	34.3	0.706	0.588	0.808

Table 8: Predictive accuracy of specifications employing one hypothetical choice variable and one non-choice rating

Model	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75				Chow Test
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)	
Benchmarks									
Myopia	−7.50	93.3	1.001	NA	−7.50	93.3	0.690	NA	NA
RPF predictor	0	37.3	0.994	-3.388	0	26.4	1.000	0.946	NA
Augmented predictors: LASSO	0.04	35.9	1.021	0.750	-0.02	26.4	1.053	0.835	NA
Hyp – preferred univariate	−0.55	36.2	0.919	0.531	−0.44	36.4	0.692	0.523	0.860
Hypothetical plus:									
Approve/disapprove	−1.05	33,8	0.964	0.662	−0.60	33.1	0.711	0.659	0.118
Happiness	−1.07	34.3	0.950	0.628	−1.48	33.5	0.722	0.725	0.142
Liking	−2.78	39.7	0.987	0.783	−4.03	47.8	0.722	1.064	0.001
Regret	−0.62	36.1	0.921	0.537	−1.05	35.8	0.701	0.573	0.061
Tempting	−1.00	35.8	0.935	0.568	−1.64	36.2	0.709	0.630	0.458
Enjoy if harmless	−1.09	35.7	0.938	0.576	−1.68	36.2	0.708	0.635	0.365
Good/bad for you	−0.63	36.1	0.922	0.538	−0.83	35.9	0.698	0.553	0.118
Pos/neg impression	−0.60	36.1	0.921	0.536	−1.36	35.8	0.705	0.602	0.077
Over/understate	−0.76	35.9	0.927	0.550	−0.91	35.8	0.699	0.560	0.211
Hyp – duplicate	0.70	33.7	0.936	0.600	0.90	34.3	0.706	0.588	0.808

Table 9: Predictive accuracy of optimized specifications

Model	Predicting from $0.75 to $0.25				Predicting from $0.25 to $0.75			
	Average bias	MSPE	Calibration (level)	Calibration (change)	Average bias	MSPE	Calibration (level)	Calibration (Δ)
Benchmarks								
Myopia	−7.50	93.3	1.001	NA	−7.50	93.3	0.690	NA
RPF predictor	0	37.3	0.994	-3.388	0	26.4	1.000	0.946
Augmented predictors: LASSO	0.04	35.9	1.021	0.750	-0.02	26.4	1.053	0.835
Hyp – preferred bivariate	0.16	32.3	0.924	0.645	−0.19	31.8	0.728	0.674
All hyp. & ratings								
LASSO	−0.17	31.6	0.951	0.752	−0.62	29.2	0.759	0.847
OLS – CV-MSPE optimized	−1.59	33.1	0.943	0.843	0.03	30.6	0.743	0.732
OLS – CV-AMPE optimized	0.87	33.8	0.882	0.639	-0.69	30.0	0.743	0.801
OLS – CV-Calib optimized	-0.10	29.8	0.931	0.814	-0.16	28.1	0.769	0.831
All hyp., LASSO	2.35	35.8	0.937	0.754	-4.10	47.4	0.752	0.681
Hyp. & all ratings, LASSO	-3.08	41.7	0.990	0.849	3.81	44.1	0.748	1.243
All hyp., ratings, & phys., LASSO	-1.93	33.6	0.972	0.968	-0.70	28.2	0.771	0.902
RPF, all hyp., ratings, & phys., LASSO	0.13	29.6	1.036	0.875	-0.04	22.5	1.050	0.873