Simple Solutions for Complex Problems in Behavioral Economics

B. Douglas Bernheim
Stanford University

April 2014
Introduction

Motivating premise:

- Existing limitations on our understanding of human behavior pose serious challenges for standard approaches to positive and normative issues in behavioral economics

Overall agenda:

- Develop methods of usefully conducting applied analyses of positive and normative issues in behavioral economics without assuming more about human behavior than is actually known
The Challenges of Positive BE

Behavior displays complex and poorly understood patterns of context dependence

Examples of context-dependence in choice:

- Anchoring (e.g., Ariely, Loewenstein, and Prelec, QJE, 2003)
- Salience (e.g., Bordalo, Gennaioli, and Shleifer, QJE, 2012)
- Priming (e.g., Chen and Shi, AER, 2009, Benjamin, Choi, and Strickland, AER, 2010)

Real economic choices display complex context-dependence

- Saez (2009) - "Details matter" (retirement saving)
- Bertrand et al. (2005) - "What's psychology worth?"
Behavioral economics rarely comes to grips with context dependence

What determines anchoring/saliency/priming, etc.? (Points to Bordalo et al., QJE, 2012, for trying)

In the theory of reference-dependent preferences, what determines reference points? (Points to Koszegi & Rabin, QJE, 2006, for trying)
Models with *reference-dependent preferences* permit a fortuitous choice of the reference point in order to fit an observed choice pattern

- Babcock, Camerer, and Loewenstein, *QJE*, 1997
- Applications involving anomalies in financial markets

There is also some direct evidence that reference points are context-dependent (Plott and Zeiler, *AER*, 2005, 2007)
One consequence: *limited ability to make robustly accurate out-of-sample predictions*

Another consequence: *vast proliferation of (simple) theories*

Examples of theory proliferation:

- Choice in the ultimatum game (fairness, intentionality, anger)
- The Allais paradox

Tests of competing theories often point in different directions

Debates over which test is "correct" and which theory is "right" may miss the point that behavior is more complex than any of these simple models because it varies by context
The Challenges of Normative BE

The problem: how should we think about an individual’s well-being when her choices are not entirely consistent?

- Suppose \( a \) is chosen over \( b \) in one setting, but \( b \) is chosen over \( a \) in another
- Suppose \( a \) is chosen over \( b \), which is chosen over \( c \), which is chosen over \( a \)
An economist’s natural instinct:

Develop understanding of *why* people make these choices through a structural model wherein choice is ultimately derived from *"true" preference*

But our theories often have many unobservable elements

The general challenge involves recoverability (Bernheim, 2009):

- To rationalize non-standard behavior, we must broaden the class of potential explanations
- To identify preferences, we must limit the class of potential explanations
- Having stepped away from the standard framework, where do we place the limits, and how do we justify them?
An Example

Positive model:

\[ C(X) = \arg \max_{x \in X} s(x) \quad s.t. \quad u(x) \geq t(X) \]

Alternative interpretations:

- \( u \) is preference, \( s \) is salience, and \( t \) is a utility threshold

- \( s \) is preference, \( u \) is salience, and \( t \) is a salience threshold
Paths forward?

Alternative methods are needed that acknowledge and embrace our partial and imperfect understanding of choice.

These methods would be *structurally minimalistic*:

- Use information about the structure of decision making processes whenever it is there is sufficiently reliable foundation for doing so, but
- Not require us to assume anything about the structure of these processes for which we lack sufficient foundation in order to conduct positive and normative economic analysis.

To be clear: better knowledge of structure may allow us to give better answers to positive and normative questions.
Normative Methods

Framework

Applications
**The problem**: how should we think about an individual’s well-being when her choices are not entirely consistent?

- Suppose $a$ is chosen over $b$ in one setting, but $b$ is chosen over $a$ in another
- Suppose $a$ is chosen over $b$, which is chosen over $c$, which is chosen over $a$

**Elements of the framework**: The Bernheim-Rangel framework is a (primarily) choice-based generalization of standard welfare economics that consists of two elements:

- *First*, a notion of *mistakes* that is potentially used to reduce the set of choices deemed welfare-relevant
- *Second*, a *welfare criterion* derived from the pattern of welfare-relevant choices.
Element #1: Mistakes

In the standard model of choice, wherein preferences are essentially synonymous with choices, there is no notion of mistakes.

In behavioral economics, normative analysis would seem to require such a notion.

But how does one define a mistake, without referencing "true preferences," which we have disavowed?
Motivating example

In some cases, mistakes seem obvious:

"American visitors to the United Kingdom suffer numerous injuries and fatalities because they often look only to the left before stepping into streets even though they know traffic approaches from the right. One cannot reasonably attribute this to the pleasure of looking left or to masochistic preferences. The pedestrian's objectives -- to cross the street safely -- are clear and the decision is plainly a mistake." - Bernheim and Rangel (2004)

Likewise, "optimal policy" (holding the pedestrian back) seems obvious

Calling this a "mistake" on the grounds that preference are "obvious" is neither objective nor generalizable; other criteria are equally problematic (e.g., relying on expressions of regret arbitrarily privileges the ex post perspectives)

Abstract illustration:

- Individual is presented with a choice between options x and y.
- In one context (frame), he chooses x over y, incorrectly thinking that y is z.
- In another context (frame), he chooses y over x, correctly recognizing that x is x and y is y.
- The combination of context-dependence and mischaracterization in the first frame involves characterization failure.
- The choice of x over y in the first frame is a "mistake" -- i.e., not a suitable guide for a policy maker who must choose between x and y on behalf of the individual.
Application to the American pedestrian in London:

- The decision to step in front of the onrushing car is (implicitly) context dependence
- Evidence could be developed to show that there was characterization failure in the naturally occurring decision frame because the pedestrian was inattentive

Substantive applications include:
- Bernheim-Rangel (2004) on addiction
- Ambhuel, Bernheim, and Lusardis (in progress) on financial education
Element #2: The welfare criterion

If choices satisfy WARP within the welfare-relevant domain, then one can proceed as if "true preferences" have been recovered.

But there is no guarantee that welfare-relevant choices will satisfy WARP.

Bernheim-Rangel (2009) identify several (arguably) desirable properties of a choice-based welfare criterion:

- Coherence (an acyclic binary relation)
- Respect for choice
- No alternative can be classified as a "mistake" based on choice patterns alone
Theorem: The only welfare criterion satisfying these properties is the *unambiguous choice relation* \( (xP^*y \text{ iff } y \text{ is chosen in no situation where } x \text{ is available}) \)

Other features of this criterion:
- Universally applicable
- Specializes to the standard welfare criterion when choices satisfy standard axioms
- Leads to counterparts for the standard tools of applied welfare analysis (compensating and equivalent variation, consumer surplus, Pareto optimality, contract curve)

Essence of approach is to exploit coherent aspects of choice while living with ambiguity implied by lack of complete coherence
An empirical application

Based on Ambhuel, Bernheim, and Lusardi (in progress), "The Welfare Effects of Financial Education"

- Illustrates the application of Element #1 (mistakes and characterization failure)

Background:

- There is some evidence that financial education impacts financial choices (Bernheim, Garrett, and Maki, 2001, Bernheim and Garrett, 2003, and others)
- Main issue: Welfare effects: What are "true preferences"? Decision skill or indoctrination/brow-beating?
- Other issues: heterogeneous interventions, heterogeneous populations, possible endogeneity, distance between educational content and measured behaviors
**Experiment**

Focus on behavioral effects of a narrow intervention involving compound interest

*Treatment group*: watches video on compound interest based on a leading popular investment book (Malkiel and Ellis)

Components of treatment:
- Simple explanation of compound interest
- Explanation and application of the "Rule of 72"
- Some rhetoric

*Control group*: Similar video on an irrelevant economic topic

*Experimental tasks*: All subjects filled out 20 "price lists," each consisting of 10 objectively identical pairs
Example:
- "Unframed choice": $X today versus $20 in 36 days
- "Framed choice": $Y today versus a $10 investment earning 2% per day for 36 days

Unframed task involves an expression of preferences concerning timing; framed task involves the same expression of preferences but also requires an understanding of compound interest.

Idea: If the choices differ within pairs, and if the evidence indicates that subjects had a poor command of compound interest, then the framed tasks evidently involve characterization failure, and the potential welfare loss is $|X - Y|$. 
Short financial literacy quiz implemented at outset

Longer test (including five questions on compounding) implemented at conclusion
  - Performance incentivized
  - Informed of incentives at outset, prior to watching video

240 subjects recruited through Amazon Mechanical Turk
  - Employed, in 20s and 30s
  - Long sessions + high pay = very low attrition
  - Unsolicited expressions of appreciation for financial education
Effects of treatment on test scores and use of rule-of-72
Effects of treatment on test scores and use of rule-of-72
Effects of treatment on choices (valuations)
Effects of treatment on choices (valuations)
So, with...

- Large increase in comprehension of compounding as indicated by test scores
- Large increase in self-reported use of rule-of-72 for decision problems
- Large increase in valuations for framed choices (with only a small degree of overshooting on average)

... we would expect large increases in welfare.
Effects of treatment on welfare
Effects of treatment on welfare
Raises question: to which part(s) of the educational intervention is behavioral responding?

Fielded 2 new treatments
- For one, removed all Rule-of-72 material
- For the other, removed all rhetoric
- For both, kept the basic explanation of compound interest
Effects on test scores

![Bar chart showing effects on test scores for different treatments. The chart compares Control, Full Treatment, Only Rule of 72, and Only Rhetoric.]
Effects on test scores
Effects on test scores
Effects on use of Rule of 72
Effects on use of Rule of 72
Effects on use of Rule of 72
Effects on choices (valuations)
Effects on choices (valuations)
Effects on choices (valuations)
Effects on welfare
Effects on welfare
Effects on welfare
Positive Methods

What is the objective of positive behavioral economics?

- One view: understand the structure of human decision making, identify the correct model of choice, etc.

- An alternative view: to predict behavior accurately in as-yet-unobserved situations

- Understanding structure is a means to an end, and not an end in itself
  - If we can make predictions accurately with imperfect knowledge of structure, we can do positive economics
  - And, using the Bernheim-Rangel framework, we can also do normative economics
Why is behavioral prediction challenging?

The complexity of context-dependence renders both non-structural and structural prediction problematic.
Motivation for the proposed approach

Reduce the alternatives that are available in diverse decision problems to vectors in some low-dimensional \textit{fundamental attribute space} (much as in Lancaster, 1966)

Estimate the relationship between fundamental attribute profiles and choices across decision problems

Apply this relationship to the fundamental attribute profiles for new decision problems to predict choices
What could we use for fundamental attributes?

Choice ultimately depends on *internal representations*, not *external characteristics*

Our perceptual and cognitive apparatus reduces complex external alternatives to comparatively simple internal representations involving *fundamental subjective attributes* (i.e., how they address a relatively small collection of basic human motivations)

*Idea*: variation that looks different in characteristic space may look similar in fundamental subjective attribute space

Example: bundles of groceries
Problem: We can't measure these fundamental subjective attributes objectively

However, we can measure them subjectively (by asking people how they perceive the alternatives available in various contexts as addressing various motivations)

As long as there is a stable relationship between these subjective measures and the actual subjective attributes, we should be able to use the former in place of the latter for purposes of predicting choices

Even if the measurement bias varies by context, it should be related to subjective aspects of the problem, for which we can proxy in a similar way
Choice:

A vs. B with

C vs. D with

E vs. F with

Average non-choice reactions:

Reaction 2

A

B

Reaction 1

36% A

64% B

C

D

Reaction 1

82% C

18% D

Reaction 2

E

F

Projected:

43% E

57% F
Candidates for fundamental subjective attributes

- Perceived effectiveness of alternatives at addressing basic motivations

- "Motivational aggregates" (hypothetical choices, liking, anticipated happiness, etc.)

- Physiological responses
Proof of Concept

- Based on Bernheim, Bjorkegren, Naecker, and Rangel (2013)

- Concrete, stylized illustration of the issues (which we implement in the lab)
  - Large number of snack items have seen sold at 25 cents (in a setting where demands are independent)
  - Task: predict demands if prices were raised to 75 cents
  - Price change “stands in” for more interesting changes in economic conditions (either standard or behavioral) – i.e., here, prices serve as the “flags”
• Analytic challenges
  ▪ Can’t use reduced-form estimation
  ▪ A structural model could use variation in weight fixing other characteristics, but predictions are very poor
  ▪ What do you do?

• Our approach
  ▪ For each choice problem (item-price pair), assess various non-choice reactions and aggregate across subjects to form measures of subjective characteristics
  ▪ Estimate relationships between purchase frequencies and subjective characteristics across choice problems with 25 cent data
  ▪ Use them along with subjective characteristics to predict purchase frequencies for 75 cent data
• Remark on model selection
  ▪ With many non-choice reactions, there will be many subjective characteristics, and hence a huge number of potential models
  ▪ Rules of the game: model selection must be based on within-estimation-sample criteria

• Bottom line: method works very well in this experimental task
  ▪ In the best specifications, we predict the average change in demand with an error of a couple of percent, and predict the variation in price sensitivity across items
  ▪ Matches or beats demanding benchmarks that use additional choice data and conventional methods
Relation to literature

- There are related strands of literature in several disciplines
- Economics (*stated preference* & *contingent valuation*)
- Marketing (especially on *stated intentions*)
- Political science (predicting elections from polls)
- Most closely related strand is on *ex post statistical calibration* in the SP literature
Distinctive aspects of our method (relative to the SP literature on ex post calibration):

1. We focus on predicting choice distributions across choice problems based on subjective characteristics of those problems; they focus on predicting individuals’ choices within a choice problem based on individuals’ characteristics, and those relationships do not port well across choice problems.

2. We avoid cross-contamination by asking real, hypothetical, & subjective questions of different people.

3. We use multiple hypothetical & subjective questions in combination; they are concerned with “correcting” one at a time.

4. We compare performance to benchmarks that use conventional data and additional choice data.
Structure of the experiment

- Two stages:
  - Computer-based choice or evaluation task with stimuli presented in random order (~30 minutes)
  - Waiting period, during which no snacks allowed unless provide according to rules of the session (~30 minutes)
- 189 snack items, 2 prices ($0.25 & $0.75)
- One group of 30 subjects made real choices
  - One implemented at random
  - Use to construct “real purchase frequencies” - RPFs
  - Verified that subjects treat these choices as “real” rather than hypothetical despite low probabilities
  - RPF falls from 27.8% at $0.25 to 20.3% at $0.75
• Groups of 28 subjects made hypothetical choices
  ▪ One elicitation protocol per group – basic, “cheap talk,” certainty scales, vicarious, and WTP
  ▪ Only difference from “real choice” group: nothing is implemented
  ▪ Used to construct “hypothetical choice frequencies” – HPFs
• Multiple groups of ≥ 28 subjects made other subjective evaluations
  ▪ Norms, happiness, liking, regret, perceived temptation, enjoyment, impact on health, impact on social image, tendency to overstate/understate inclination to each
  ▪ Used to construct “subjective response frequencies”
Prediction task and evaluation criteria

- Reminder of task:
  - You observe demand (RPFs) for all items at a single price (either $0.25 or $0.75)
  - Object: estimate the change in demand for each item if its price were changed to the other value
  - Note: you observe no price variation either within or across items
  - Price stands in for any economic condition, the effect of which is difficult to measure with standard methods
• Performance metrics
  ▪ Mean prediction error, MPE (“average bias”)
  ▪ Mean-squared prediction error, MPSE (overall accuracy)
  ▪ Calibration (bias conditional on a given prediction)
    Measure of calibration: value of $\beta$ in the regression
    \[
    \text{Actual} = \alpha + \beta \cdot \text{Predicted}
    \]
    Perfect calibration is $\beta = 1$ (so that expected value changes unit for unit with the prediction)
    We report calibration for predicted levels, and for predicted changes in RPF
Benchmarks

- Benchmarks using the same limited choice data as our method
  - Tried a “natural” structure model that exploits variation in weight of items, controlling for other characteristics, but performed terribly

- Benchmarks using additional choice data (RPFs at target price for other items)
  - Regressions of RPF at target price on RPF at observed price, by itself or with controls for items’ characteristics (OLS and LASSO)
  - LASSO specification performed best
## Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>LASSO predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
</tr>
<tr>
<td>Calib, levels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 to 25</td>
<td>1.001</td>
<td>1.021</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
</tr>
<tr>
<td>Calib, diffs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 to 25</td>
<td>NA</td>
<td>0.780</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
</tr>
</tbody>
</table>
Performance of HPFs as predictions

- Generally, results confirm what we know from the literature:
  - Hypothetical answers generally overstate inclination to pay by a significant margin (6.9 percentage points)
  - Still, HPFs are strongly correlated with RPFs
- In our prediction task, they underperform the myopic benchmark
- Evidence on alternative protocols is a bit mixed, but generally they don’t do much better
## HPFs as Predictions

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>HPF, std protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>7.39</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-7.39</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>137.4</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>137.4</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>0.610</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>0.500</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.248</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.248</td>
</tr>
</tbody>
</table>
Performance of predictive models

- A large number of subjective attributes can be used as predictors
  - Variables based on all hypothetical and subjective responses
  - Quadratic terms and interactions

- Critical to select models using within-sample criteria

- For models with large numbers of potential predictors, we use:
  - LASSO
  - Maximize a performance metric in cross-validation
## Optimized Predictions (LASSO)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Small bias: 2 to 8%
## Optimized Predictions (LASSO)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
</tr>
</tbody>
</table>

On average, MSPE better than toughest benchmark
### Optimized Predictions (LASSO)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Calibration for levels nearing toughest benchmark.
## Optimized Predictions (LASSO)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Calibration for diffs roughly matches toughest benchmark
Another possibility is to select models by maximizing within-sample cross-validated performance.

- Natural choice for the objective is to minimize cross-validated MSPE, absolute MPE, or distance of the calibration parameter from unity.
- We use a standard hill-climbing algorithm, starting from the LASSO model.
- Optimizing calibration within-sample seems to work especially well for our-of-sample predictions.
# Optimized Predictions (CV)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
<th>Optimize CV calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg bias:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
<td><strong>-0.10</strong></td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
<td><strong>-0.16</strong></td>
</tr>
<tr>
<td><strong>MSPE:</strong> 75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
<td>29.8</td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
<td>28.1</td>
</tr>
<tr>
<td><strong>Calib, levels</strong> 75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
<td>0.931</td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
<td>0.769</td>
</tr>
<tr>
<td><strong>Calib, diffs</strong> 75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
<td>0.814</td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
<td>0.831</td>
</tr>
</tbody>
</table>

**Trivial bias:** 1% to 2%
# Optimized Predictions (CV)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
<th>Optimize CV calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias:</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
<td>-0.10</td>
</tr>
<tr>
<td>75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
<td>-0.16</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
<td><strong>29.8</strong></td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
<td><strong>28.1</strong></td>
</tr>
<tr>
<td>Calib, levels</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
<td>0.931</td>
</tr>
<tr>
<td>75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
<td>0.769</td>
</tr>
<tr>
<td>Calib, diffs</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
<td>0.814</td>
</tr>
<tr>
<td>75 to 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
<td>0.831</td>
</tr>
</tbody>
</table>

Beats everything on MSPE
## Optimized Predictions (CV)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
<th>Optimize CV calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
<td>-0.10</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
<td>-0.16</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
<td>29.8</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
<td>28.1</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
<td><strong>0.931</strong></td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
<td><strong>0.769</strong></td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
<td><strong>0.814</strong></td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
<td><strong>0.831</strong></td>
</tr>
</tbody>
</table>

No further improvement in calibration for levels
## Optimized Predictions (CV)

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>LASSO</th>
<th>Optimize CV calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.17</td>
<td>-0.10</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>-0.62</td>
<td>-0.16</td>
</tr>
<tr>
<td>MSPE: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>31.6</td>
<td>29.8</td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td>29.2</td>
<td>28.1</td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.951</td>
<td>0.931</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.759</td>
<td>0.769</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.752</td>
<td>0.814</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.847</td>
<td>0.831</td>
</tr>
</tbody>
</table>

Impressive calibration for differences
• Question: do we really need all of these predictors?

• Consider univariate and bivariate models, selecting the best ones through AIC

• The HPF variables tend to be selected, and to result in the best predictions

• Among those, the standard protocol HPF tends to be selected first, and to result in the best predictions

• Performance isn't bad, but not as good as the full models, particularly with respect to calibration in differences
# Univariate and Bivariate Predictions

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>Selected univariate model</th>
<th>Selected bivariate model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg bias:</strong> 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.55</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>25 to 75</td>
<td>-0.02</td>
<td>0.44</td>
<td>-0.19</td>
</tr>
<tr>
<td><strong>MSPE:</strong> 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td>36.2</td>
<td>32.3</td>
</tr>
<tr>
<td></td>
<td>25 to 75</td>
<td>26.4</td>
<td>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td><strong>Calib, levels:</strong> 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.919</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>25 to 75</td>
<td>1.053</td>
<td>0.692</td>
<td>0.728</td>
</tr>
<tr>
<td><strong>Calib, diffs /5 to 25</strong></td>
<td>NA</td>
<td>0.780</td>
<td>0.531</td>
<td>0.645</td>
</tr>
<tr>
<td></td>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.674</td>
</tr>
</tbody>
</table>
### Univariate and Bivariate Predictions

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>Selected univariate model</th>
<th>Selected bivariate model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg bias: 75 to 25</td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.55</td>
<td>0.16</td>
</tr>
<tr>
<td>25 to 75</td>
<td>7.50</td>
<td>-0.02</td>
<td>0.44</td>
<td>-0.19</td>
</tr>
<tr>
<td>MSPF: 75 to 25</td>
<td>93.3</td>
<td>35.9</td>
<td><strong>.36.2</strong></td>
<td><strong>32.3</strong></td>
</tr>
<tr>
<td>25 to 75</td>
<td>93.3</td>
<td>26.4</td>
<td><strong>36.4</strong></td>
<td><strong>31.8</strong></td>
</tr>
<tr>
<td>Calib, levels 75 to 25</td>
<td>1.001</td>
<td>1.021</td>
<td>0.919</td>
<td>0.924</td>
</tr>
<tr>
<td>25 to 75</td>
<td>0.690</td>
<td>1.053</td>
<td>0.692</td>
<td>0.728</td>
</tr>
<tr>
<td>Calib, diffs 75 to 25</td>
<td>NA</td>
<td>0.780</td>
<td>0.531</td>
<td>0.645</td>
</tr>
<tr>
<td>25 to 75</td>
<td>NA</td>
<td>0.835</td>
<td>0.523</td>
<td>0.674</td>
</tr>
</tbody>
</table>

On average, MSPE close to toughest benchmark
## Univariate and Bivariate Predictions

<table>
<thead>
<tr>
<th></th>
<th>Myopic</th>
<th>Best benchmark</th>
<th>Selected univariate model</th>
<th>Selected bivariate model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg bias: 75 to 25</strong></td>
<td>-7.50</td>
<td>0.04</td>
<td>-0.55</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>25 to 75</strong></td>
<td>7.50</td>
<td>-0.02</td>
<td>0.44</td>
<td>-0.19</td>
</tr>
<tr>
<td><strong>MSPE: 75 to 25</strong></td>
<td>93.3</td>
<td>35.9</td>
<td>36.2</td>
<td>32.3</td>
</tr>
<tr>
<td><strong>25 to 75</strong></td>
<td>93.3</td>
<td>26.4</td>
<td>36.4</td>
<td>31.8</td>
</tr>
<tr>
<td><strong>Calib, levels 75 to 25</strong></td>
<td>1.001</td>
<td>1.021</td>
<td><strong>0.919</strong></td>
<td><strong>0.924</strong></td>
</tr>
<tr>
<td><strong>25 to 75</strong></td>
<td>0.690</td>
<td>1.053</td>
<td><strong>0.692</strong></td>
<td><strong>0.728</strong></td>
</tr>
<tr>
<td><strong>Calib, diffs 75 to 25</strong></td>
<td>NA</td>
<td>0.780</td>
<td><strong>0.531</strong></td>
<td><strong>0.645</strong></td>
</tr>
<tr>
<td><strong>25 to 75</strong></td>
<td>NA</td>
<td>0.835</td>
<td><strong>0.523</strong></td>
<td><strong>0.674</strong></td>
</tr>
</tbody>
</table>

*Calibration OK, but weak in differences*
• **Bottom line:**
  - Based on subjective characteristics of choice problems, we can predict how choice frequencies respond to interventions with no prior parallels
  - Accuracy compares favorably to predictions based on standard methods that require observation of parallel interventions
Directions for further work

- Continue search for “universal type space” by exploring other subjective characteristics, including biometrics
- Refine within-sample model selection
- Explore performance for other interventions, in other choice domains, and across domains
- Attempt applications
Assessing the potential value of biometric measurements

- Based on Smith, Bernheim, Camerer, and Rangel (2013)

- Experiment:
  - 17 subjects view 100 snacks while undergoing full brain scans (fMRI)
  - After the scan, they make choices over 50 randomized pairs, with one choice implemented at random
  - Critical: choices are not anticipated when scans are conducted (so these are non-choice reactions)
• Analysis
  ▪ For each subject, 48 prediction models are then calibrated, each based on a sample that omits two choices
  ▪ Models relate choices to differential brain activity
  ▪ Attention to variable selection and overfitting problems is required
  ▪ Predicted choices are then compared to actual choices
• Predictions prove reasonably accurate for just over half of the subjects, and not at all for the others
Out-of-Sample Predictive Accuracy

A

B
Concluding Remarks