An Approach to Testing Reference Points

Alex Rees-Jones and Ao Wang*

November 30, 2022

Abstract

We present a general approach to experimentally testing candidate reference points. This approach builds from Prospect Theory's prediction that an increase in payoffs is perfectly offset by an equivalent increase in the reference point. Violation of this prediction can be tested with modifications to existing econometric techniques in experiments of a particular design. The resulting approach to testing theories of the reference point is minimally parametric, robust to broad classes of heterogeneity, yet still implementable in comparatively small sample sizes. We demonstrate the application of this approach in an experiment that tests the role of salience in setting reference points.

Keywords: reference points, prospect theory, nonparametric econometrics.
JEL Codes: C14, D9.

*Rees-Jones: University of Pennsylvania and NBER, email: alre@wharton.upenn.edu. Wang: National University of Singapore, email: ao.wang@nus.edu.sg. We are grateful to Ned Augenblick, Stefano DellaVigna, Lorenz Goette, Ori Heffetz, Alex Imas, Botond Kőszegi, Francesca Molinari, Muriel Niederle, Ted O'Donoghue, Collin Raymond, and Charlie Sprenger, as well as seminar audiences at the BEEMA Annual Conference, the ESA Annual Meetings, the National University of Singapore, SITE–Experimental Economics, Stanford University, the SOBE Conference, Swarthmore College, UC Berkeley, UC Santa Barbara, and the University of Pennsylvania for helpful advice on this project. We thank the University Research Foundation at the University of Pennsylvania and the Wharton Behavioral Lab for financial support. The project described in this paper relies on data from surveys administered by the Understanding America Study, which is maintained by the Center for Economic and Social Research (CESR) at the University of Southern California. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the USC, the UAS, or any other entity.

Prospect Theory (Kahneman & Tversky, 1979) is the most prominent theoretical framework advanced by behavioral economists. A central component of prospect theory is reference dependence: the idea that, rather than evaluating final wealth or consumption levels, individuals instead evaluate gains and losses relative to a reference point. Prospect Theory also incorporates a number of additional components regarding how gains and losses are evaluated (i.e., with loss aversion and diminishing sensitivity), how probabilistic outcomes are mentally represented (i.e., with combination, segregation, or cancellation), and how probabilities are ultimately assessed (i.e., through the probability weighting function). In economic applications of Prospect Theory, however, there is significant variability in whether, and how, these additional assumptions are imposed,[1] and active dispute about the importance of some of them.[2] Reference-dependence is viewed by many as the most supported and agreed-upon component of Prospect Theory, and it accordingly has been widely adopted in economic applications.

While the idea that individuals consider gains and losses is widely accepted, the coding of gains and losses remains imperfectly understood and actively disputed. To define a gain or a loss, one must define the reference point. Kahneman and Tversky themselves viewed the reference point as an object that could change across contexts and potentially be manipulated.[3] Supporting this view, the behavioral economics literature contains studies advancing a wide variety of candidate reference points, including current asset position (Kahneman & Tversky, 1979), endowments (Kahneman et al., 1990), social comparisons (Schwerter, 2013), goals (Heath et al., 1999), targets (Pope & Schweitzer, 2010), averages (Crawford & Meng, 2011), adaptive averages (Thakral & Tô, 2021), and expected distributions of consumption

[1] For example, many applications study loss aversion in isolation while assuming away diminishing sensitivity, while others rely critically on diminishing sensitivity alone. Formal attention and modeling to the combination, segregation, or cancellation assumptions is rare. Probability weighting is assumed away in many Prospect Theory papers, and yet is viewed as essential in some others.

[2] Implicit dispute about the central importance of probability weighting or diminishing sensitivity can be inferred by how often these features are left out of models building on prospect theory. More explicit dispute about the central importance of loss aversion has recently received significant attention. For example, Gal & Rucker (2018) argues that current evidence does not support the foundational claim that losses are more impactful than gains, and Chapman et al. (2018) argues that loss-loving behavior is prevalent, and possibly even as prevalent as loss aversion.

[3] "However, the location of the reference point, and the consequent coding of outcomes as gains and losses, can be affected by the formulation of the offered prospects and by the expectations of the decision-maker." (Kahneman & Tversky, 1979, page 247)

(Kőszegi & Rabin, 2006).[4] In our view, this literature suggests that many of these candidates may be adopted in particular environments, and thus that none of them is "the" reference point at all times. Thus, for Prospect Theory to be fully specified, behavioral economists must develop and empirically validate a relatively nuanced theory of the reference point: one that dictates which reference points are adopted in different situations.

Pursuing a nuanced theory of the reference point is conceptually straightforward but challenging to implement. It is conceptually straightforward because it can be naturally pursued with the standard scientific method: theorizing that certain factors influence reference point adoption, randomly varying those factors, testing if they do indeed set the reference point as theorized, updating theories, and then repeating the process. The challenge comes in testing if reference points are indeed set as theorized. Reference points are typically viewed as unobservable, and thus their value must be inferred from choice behavior. To the extent that a general and portable method for this inference currently exists, it is arguably the bunching-based methodology advanced in papers like Allen et al. (2017), Rees-Jones (2018), and Seibold (2021). This approach relies on the sharp change in marginal utilities that occurs at the reference point when loss aversion is assumed, which in certain classes of models generates excess mass at that point. While seeing excess mass at a particular value provides compelling evidence of the precise reference point that is adopted, formal statistical testing for the presence of excess mass relies on very large sample sizes. These data requirements make this approach a difficult foundation for designing a long sequence of tailored experiments in the scientific process described above, with the approach instead finding most use in large-scale datasets of convenience. Additionally, this approach relies critically on auxiliary assumptions beyond mere reference dependence—for example, regarding the functional form of loss aversion. A researcher who "buys in" to the idea of reference dependence, but remains unsure about the specification of these additional components of prospect theory, may thus view such approaches with skepticism. While there are a wide variety of other approaches to testing specific candidate reference points presented in the literature, we are not aware of one that avoids all of these concerns.

[4]While we cite a single focal paper as an example of each style of reference point, for most of these candidates there are many papers available. See Brown et al. (2020) for a metaanalysis of Prospect Theory (specifically focusing on loss aversion) that summarizes 607 empirical applications.

In this paper, we propose a new approach to testing theories of the reference point. Our approach involves the combination of non-parametric econometric techniques previously unused in the lab-experimental literature and novel experimental designs that make the application of such techniques possible. This combination results in a principled hypothesis-testing framework that disentangles the assumption of reference dependence from the many other components of prospect theory discussed above. Unlike prevailing methods, this approach entirely avoids reliance on Prospect Theory's auxiliary functional form assumptions and remains valid and easily implementable even in the presence of broad classes of heterogeneity. Perhaps surprisingly, these statistical benefits arise even in comparatively small sample sizes. In short, we believe this approach provides the needed tool to work towards a theory of the reference point through the iterative scientific process.

In section 1, we describe the intuition behind our test. In our view, the defining characteristic of a reference-dependent model is that the argument of the utility function is relative rather than absolute. Loosely speaking, rather than assuming utility takes the functional form $u(c)$, reference dependent models take the form $u(c-r)$.[5] We show that the mere assumption that $c-r$ is the relevant input to utility allows for strong tests of correct reference-point specification. Given exogenous variation in both c and r, the *level sets* of utility in $c \times r$ space take on a particular form: they are parallel lines of slope 1. Put simply, for any given consumption/referent combination, the consequences of increasing consumption by one unit are offset by increasing the referent by one unit. This implies that, if experimental subjects are presented with choices between pairs of gambles where all payoffs are shifted up by a common amount (denoted Δ), this too will be offset by increasing the referent by the same amount. The probability of choosing a particular one of the gambles may therefore be described by a single-index function $g(\Delta - r)$—i.e., a function with the slope-1-parallel-line structure just described. This suggests an immediate strategy for robust inference regarding candidate reference points: randomly varying a uniform payoff shock Δ while also varying the candidate reference point, and then rejecting the reference point if choice probabilities do not admit the necessary representation.

[5]Note that some models adopt a hybrid notion of reference dependence incorporating both an absolute and a relative term. While we will often discuss intuition in the context of a purely relative model, our formal approach accommodates utility influenced by absolute and relative components.

In Section 2, we describe the formal statistical framework we adopt in order to test these predictions. The theory just discussed implies that testing for a correctly specified reference point may be done by testing for a particular single-index representation. We may thus proceed by modifying existing, powerful non-parametric techniques for single-index specification testing. Conceptually, the approach may be understood as comparing the best-fit single-index function $\hat{g}(\Delta - r)$ and the best-fit unrestricted function $\hat{g}(\Delta, r)$, each using using kernel methods, and then examining whether the differences between them is sufficiently large to reject the null of the single-index representation. We build on the work of Fan & Li (1996) to derive an analytical formula for the associated p-value, with modifications to accommodate some additional structure imposed by our model and the clustering issues that arise in our domain.

In Section 3 we demonstrate how to apply these techniques in an experiment. We begin with a discussion of general experimental design considerations. This discussion provides guidance on how to design gambles that serve best as experimental stimuli, guidance on how to vary reference points, and guidance on assessing statistical power. We then demonstrate the application of these considerations in an experiment that we deployed among 1,001 members of the Understanding America Study. In this experiment, subjects chose between a sure option and a 50-50 gamble, with payoffs from all choices facing uniform shocks of the nature required by our test. Subjects are also presented with randomized values of two variables that have been used as reference points in prior literature: average earnings of a comparison group and goals. Based on our reading of existing literature, we believed each candidate could be adopted as reference points in at least some situations.[6] Within the experiment, we sought to vary whether each reference point would be extremely salient (in which case we would expect it to be adopted) or extremely subtlety presented (in which case we would not expect it to be adopted). To make a potential referent salient, it was vividly presented in large red font over every decision that was made. In contrast, when it was not salient, the referent was not presented to subjects again after a brief mention in the

[6]Supporting averages or expectations, see for example Abeler et al. (2011), Crawford & Meng (2011), Gill & Prowse (2012), Marzilli Ericson & Fuster (2011), Schwerter (2013), or Lindskog et al. (2022). Supporting goals, see for example Heath et al. (1999), Hsiaw (2013), Allen et al. (2017), Markle et al. (2018), or Hsiaw (2018).

introductory materials.

Using these data and our testing approach, we confirm a claim that has been well documented in prior literature: that salient goals can serve as reference points. Emphasizing the importance of salience, however, our approach strongly rejects the adoption of goals as reference points when they are not made salient. We believe that this comparison of results provides some reassurance that the test is working as expected: our test appears to confirm goals work as reference points in a situation where they would be most expected to do so, but our test rejects that goals operate as reference points in a situation where their adoption appears unlikely due to experimental design. We additionally view these results as a demonstration of one of the foundational ideas motivating the pursuit of a more nuanced theory of the reference point: that the adoption of a particular candidate reference point can be influenced by changing features of the decision environment.

Perhaps more surprising results arise when examining tests of average earnings as a reference point. In either salience condition, this reference point is fails our test. These results suggest that further scrutiny of the average-earnings-based reference points in the literature may be warranted. Comparing our decision environment to those in which average-earnings-based reference points appear to be adopted can help in generating new hypotheses for factors that matter in reference point adoption.

In Section 4, we conclude. We discuss several strengths and weaknesses of our test, provide guidance on its practical usage, and discuss the implications of our experimental results for a theory of reference point formation. We also highlight further experiments that we view as necessary in the path towards refining that theory, and highlight how our techniques can be used in those future works.

Our paper contributes to a small but growing literature aiming to develop econometric techniques specifically optimized for behavioral models (see, e.g., Barseghyan et al., 2013; Strack & Taubinsky, 2021). Due in part to the history of small sample sizes in behavioral economics experiments, as well as this research community's preference for transparent reduced-form tests of comparative statics, behavioral economists have minimally engaged with theoretical econometrics. With the simultaneous rapid rise in experimental sample sizes afforded by online platforms and the rapid proliferation of structural econometrics

among behavioral economists (DellaVigna, 2018), the potential value of rectifying this blind spot in the literature has become more clear. This paper demonstrates this value in a particular salient way. Econometricians developed a large and successful technical literature on the non-parametric estimation of single-index models (see, e.g., Ichimura, 1993; Fan & Li, 1996; Horowitz, 2001; Horowitz & Mammen, 2004, 2010), but few field applications of these techniques currently exist in the literature. As we document, with some modification this body of work can be used to develop a broadly portable and easily implementable testing framework for one of the most core questions in behavioral economics.

1 A Conceptual Approach to Testing Reference Points

In this section, we begin with an intuitive discussion of the nature of our approach to testing a candidate reference point. We formalize our approach more precisely in Section 1.2.

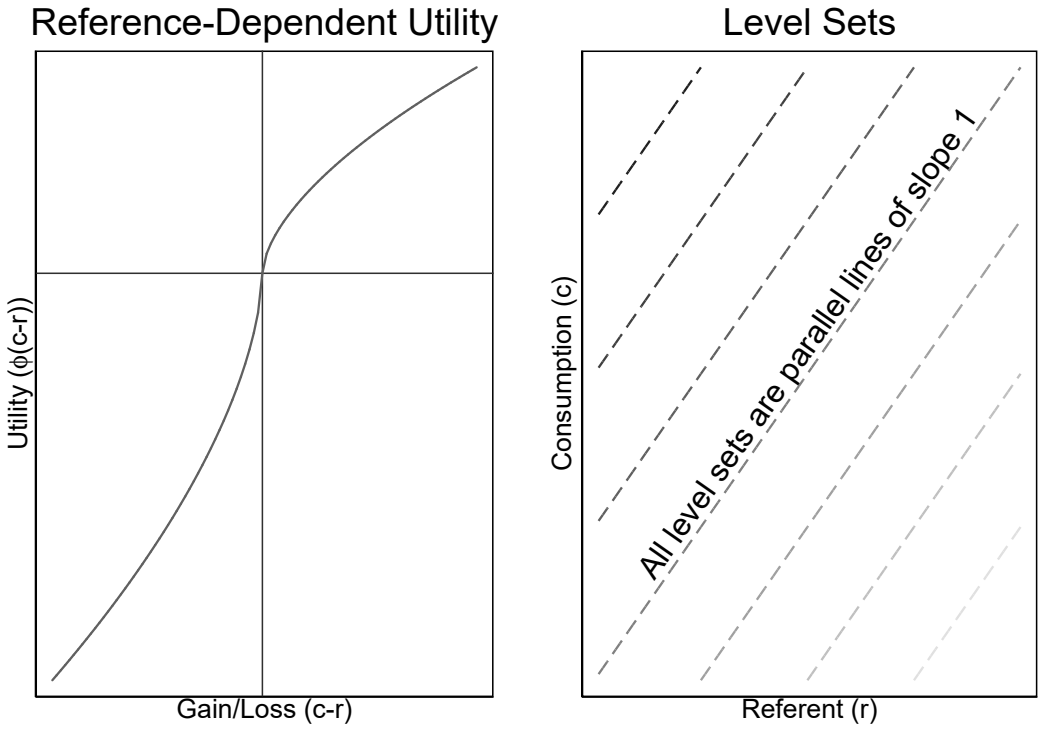
Our goal is to understand the key sources of identification when examining a model that is reference dependent, ideally relying on few other substantive assumptions. We adopt a specific and reasonably broad definition of the fundamental content of the assumption of reference dependence. Beyond technical assumptions, the core substantive assumption we wish to make is that the observed utility input $c \in \mathbb{R}$ and utility $u \in \mathbb{R}$ satisfy $u = \phi(c - r)$, where $r \in \mathbb{R}$ is a reference point (or "referent") and ϕ is a monotone function.

1.1 Intuition in Case with Direct Observation of Utility

Most economic applications treat utility as fundamentally unobservable, and thus assume that u is unobserved in the notation above. In this section, we will build towards characterizing the intuition for identification in the latent utility case, but will begin by considering the path forward in the simpler case where u is directly observed.

In such a case, the consequences of correct specification of the reference point can be completely characterized by the implications for properly defined *level sets* of u. While we are only imposing minimal functional form restrictions on ϕ—constraining the nature of its relative input and assuming that it is monotone—these core assumptions are enough to make very stark predictions about the nature of these level sets. An immediate implication of the

Figure 1: Level-Sets for Reference-Dependent Utility Function



Notes: This figure represents the level sets in $c \times r$ space for a reference-dependent utility function of the form $\phi(c-r)$. The dashed lines indicate example level sets plotted for this function, with darker lines denoting higher utility evaluations. At any potential consumption/referent combination, increasing consumption and the referent by equal amounts leads to the same gain/loss evaluation, resulting in a utility evaluation on the same level set. This generates the distinctive pattern of all level sets being parallel lines of slope 1—the key property that we examine in our test.

monotonicity of ϕ is that it is invertible, thus allowing us to express the equation as

$$\phi^{-1}(u) = c - r \tag{1}$$

$$\rightarrow \quad c = \phi^{-1}(u) + r \tag{2}$$

Put simply, in this model, every level set is a line of slope 1 in $c \times r$ space. This is visually represented in Figure 1.

This mathematical statement aligns with a simple intuition about relative thinking. If our utility of consumption is evaluated purely by relative position as compared to a referent,

then any increase in consumption can be offset by an increase in the referent of the same size. Consuming one unit compared to a referent of zero, or two units compared to a referent of one, or three units compared to a referent two (and so on) all will be evaluated as a gain of one. The assumption that the gain of 1 is all that matters for utility provides remarkable power for identification, in that it makes the stark and easily testable prediction that all levels sets are merely parallel lines of a particular slope. A violation of this property provides a basis for firmly establishing that, if utility is indeed reference dependent according to structure $u = \phi(c - r)$ for *some* reference point, then the reference point considered must be the wrong one.

1.2 Extending Intuition to Latent Utility Rationalizing Gamble Choice

The intuition above demonstrates that reference points can be tested under quite minimal assumptions when utility itself is observed and when utility takes a particularly simple reference-dependent structure. In this subsection, we demonstrate that similar results can be generated when latent and more sophisticated reference-dependent utility models rationalize binary choices, although some care is needed in the construction of that environment.

Assume the reference-dependent agent is deciding between gambles over fixed, finite sets of mutually exclusive outcomes. Each gamble $\mathcal{G} = (p_o, c_o)_{o \in O}$ is defined by the probability $p_o \in [0, 1]$ and the consumption $c_o \in \mathbb{R}$ yielded by each possible outcome $o \in O$. To avoid confusion when dealing with multiple gambles, we will denote different gambles with subscripts and will link p_o, c_o, and O with their associated gambles using superscripts.

The following assumption formalizes our assumed structure of reference-dependent choices over these gambles.

Assumption 1. ***Reference-Dependent Utility:*** *Consider two gambles, $\mathcal{G}_0$ and $\mathcal{G}_1$. $\mathcal{G}_1$ is chosen over $\mathcal{G}_0$ only if $U(\mathcal{G}_1|r) \geq U(\mathcal{G}_0|r)$, where*

$$U(\mathcal{G}|r) = \sum_{o \in O^{\mathcal{G}}} p_o^{\mathcal{G}} \cdot (\psi(c_o^{\mathcal{G}}) + \phi(c_o^{\mathcal{G}} - r)) + \epsilon. \tag{3}$$

Both ψ and ϕ are strictly increasing functions. ψ captures a standard direct utility function.

ϕ captures a reference-dependent utility function, and is assumed to be nonlinear.[7] *ϵ is the realization of an i.i.d random-utility term distributed according to the probability density function f_ϵ.*

When decisions are made according to Assumption 1, a simple single-index representation will not generally be available. However, if a specific structure is imposed on the gambles presented, such a representation can arise. When discussing the intuition of our approach in the simplified setting of Section 1.1, we noted that if gains and losses are all that matters, an increase in consumption is completely offset by an increase in the reference point of the same size. When considering gamble choices, note that we no longer have a single consumption variable, but instead have consumption amounts associated with each outcome in each gamble. In this situation, the analogous line of reasoning is that increasing consumption *across all outcomes* by a fixed amount is completely offset by an increase in the reference point of the same size. To pursue a similar test in this environment, it is thus necessary to consider sets of gamble choices in which all consumption amounts are varied in unison rather than independently varied.

Given a shifting parameter $\Delta \in \mathbb{R}$ and a base gamble $\mathcal{G}$, define the Δ-*shifted gamble* as $S(\Delta|\mathcal{G}) = (p_o^{\mathcal{G}}, c_o^{\mathcal{G}} + \Delta)_{o \in O^{\mathcal{G}}}$. Consider the behavior that would arise when subjects are presented with binary choices between $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$ for fixed base gambles $\mathcal{G}_0$ and $\mathcal{G}_1$ and a varying shifting parameter Δ. Define a variable Y to be equal to 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. It holds that

$$\begin{aligned} E[Y|\Delta, r] = & Pr\Big(\sum_{o \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} \cdot (\psi(c_o^{\mathcal{G}_1} + \Delta) + \phi(c_o^{\mathcal{G}_1} + \Delta - r)) - \\ & \sum_{o \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} \cdot (\psi(c_o^{\mathcal{G}_0} + \Delta) + \phi(c_o^{\mathcal{G}_0} + \Delta - r)) \\ & \geq \epsilon_0 - \epsilon_1 \Big) \end{aligned} \tag{4}$$

Note that the reference-dependent components can be consolidated into *a single-index function*—that is, a function with a single unidimensional input. This function is denoted by ν, and is guaranteed to be nonconstant due to the assumption that ϕ is nonlinear. Using

[7] Note that if ϕ is linear, then r serves as a choice-irrelevant constant.

this consolidating term, equation 4 can be expressed as

$$\begin{aligned} E[Y|\Delta, r] =& Pr(\nu(\Delta - r) + \\ & \sum_{o \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} \cdot (\psi(c_o^{\mathcal{G}_1} + \Delta)) - \sum_{o \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} \cdot (\psi(c_o^{\mathcal{G}_0} + \Delta)) \\ & \geq \epsilon_0 - \epsilon_1) \end{aligned} \tag{5}$$

This structure remains more complicated than the single-index representation derived in Section 1.1, but can be simplified with an additional common assumption:

Assumption 2. ***Local Linearity of Direct Utility:*** *Over the support of* Δ*,*

$$\sum_{o \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} \cdot \psi(c_o^{\mathcal{G}_1} + \Delta) - \sum_{o \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} \cdot \psi(c_o^{\mathcal{G}_0} + \Delta) = k$$

for a some constant k*.*

In words, this assumption states that the change in direct consumption utility (ψ) from adding Δ to all outcomes of base gamble $\mathcal{G}_1$ or to all outcomes of base gamble $\mathcal{G}_0$ is equal. This is guaranteed to hold exactly if consumption utility (ψ) is linear. Less restrictively, this property will be approximated when consumption utility (ψ) is approximately locally linear over a region defined by the base level of consumption and the support of Δ. Note that in circumstances where the support of Δ is narrow, this holds in common economic models. Concretely, arguments like that of the Rabin Calibration Theorem (Rabin, 2000) suggest that the curvature of the utility function is negligible over several-dollar-wide windows of wealth.

If Assumption 2 holds, this results in a final representation of

$$E[Y|\Delta, r] = Pr(\nu(\Delta - r) + k \geq \epsilon_0 - \epsilon_1) = g(\Delta - r) \tag{6}$$

for some function g. Equation 6 indicates that the conditional expectation of Y can be represented by a single-index function. To make use of this finding, we will consider it's implications for the conditional expectation that would arise when conditioning on Δ and a

candidate reference point r^c. We will first consider the case where this candidate is the true reference point, captured in the following assumption.

Assumption 3. ***Null Hypothesis:*** *$r = r^c$, where r is the true reference point applied in the utility function of Assumption 1 and r^c is a proposed candidate reference point.*

We now arrive at the Proposition that justifies our testing approach.

Proposition 1. ***Single-Index Representation for Correctly Specified Referent:*** *Consider choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. Assume that choices are governed by reference-dependent utility (Assumption 1) with locally linear direct utility (Assumption 2). Let r^c be a correctly specified candidate reference point (Assumption 3). Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. There exists a non-constant function g such that $E[Y|\Delta, r^c] = g(\Delta - r^c)$.*

In short, Proposition 1 guarantees that, under the null hypothesis, the same single-index structure that arose over (c, r) in the direct utility case arises over (Δ, r^c) in the latent utility case. This proposition follows immediately from the calculations leading up to equation 6.

1.2.1 Illustrating Levels Sets in an Example

To help illustrate the level-set structure that arises in the latent-utility case, Figure 2 presents a simple example.

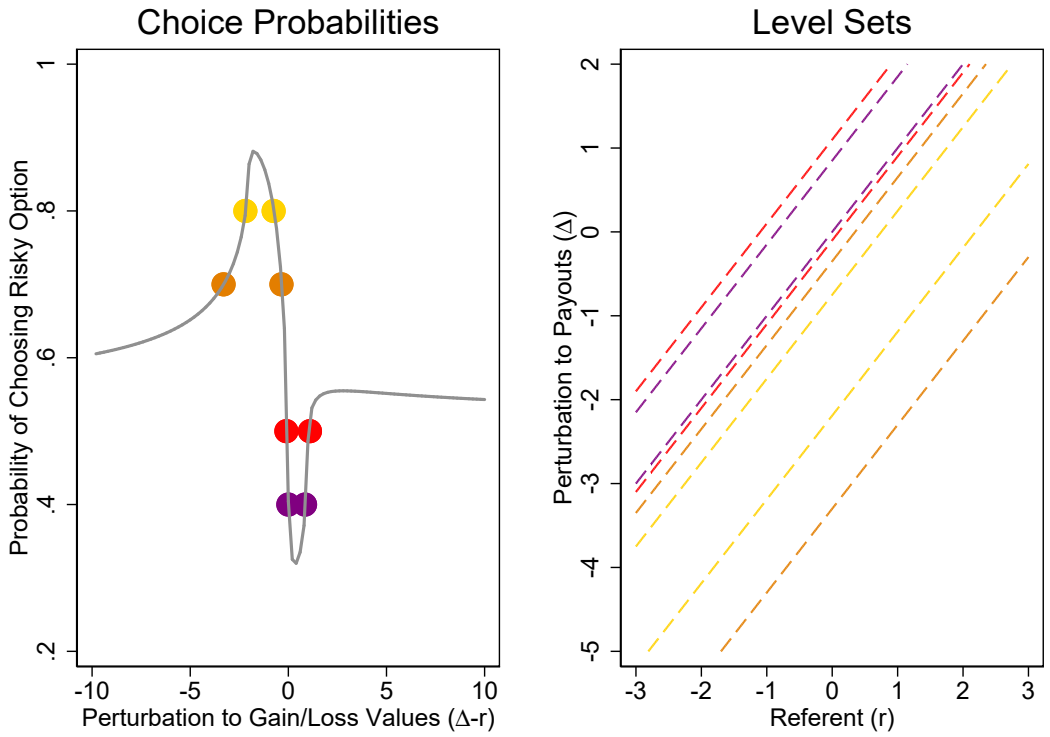
To construct this figure, we consider a situation with two base gambles: a "safe option" ($\mathcal{G}_0$) offering \$0 and "risky option" ($\mathcal{G}_1$) consisting of a 50-50 chance of +\$2 or -\$1. Our simulated individual will face choices between Δ-shifted versions of these gambles. When choosing between Δ-shifted versions of these gambles, the safe option will have the payoff $\$0 + \Delta$ and the risky option will have a 50-50 chance of $\$2 + \Delta$ or $-\$1 + \Delta$.

We assume the individual adopts a standard prospect-theory value function:

$$\phi(c|r) = \begin{cases} (c-r)^\alpha & \text{if } c \geq r \\ -\lambda(r-c)^\alpha & \text{if } c < r \end{cases}. \tag{7}$$

We assume that $\lambda = 2$ (reflecting loss aversion) and that $\alpha = 0.6$ (reflecting diminishing sensitivity). As above, assume the individual chooses the risky option only if $U(\mathcal{G}_1|r) \geq$

Figure 2: Level-Sets for Choice Probabilities with Latent Reference Dependence



Notes: This figure presents an example of choice-probability function and its level sets. This example is generated from a simulation using the utility function expressed in equation 7 and setting $\lambda = 2$ and $\alpha = 0.6$. Dots of the same color in the figure on the left map to the same choice probability for four arbitrary values. The dashed lines of the same color in the figure on the right indicate the (Δ, r) combinations that map to the color-coded choice probability value.

$U(\mathcal{G}_0|r)$, which implies $.5\phi(\$2 + \Delta|r) + .5\phi(-\$1 + \Delta|r) - \phi(\$0 + \Delta|r) \geq \epsilon_0 - \epsilon_1$. We assume that $\epsilon_0 - \epsilon_1$ is drawn from the standard normal distribution.

The left panel of Figure 2 presents the function $E[Y|\Delta, r]$, which was previously written in generality in equation 6 and which is now plotted with this specific assumed utility structure. As this function demonstrates, the probability of choosing the risky option varies substantially as Δ is varied in the vicinity of the reference point.

The right panel of Figure 2 plots the level sets associated with the colored dots from the left panel. Because the function plotted in the left panel is not monotone, multiple points on its x-axis can map to the same value on the y-axis. The dots in the left-panel figure represent specific points mapping into the level sets in the right panel of corresponding color. As in

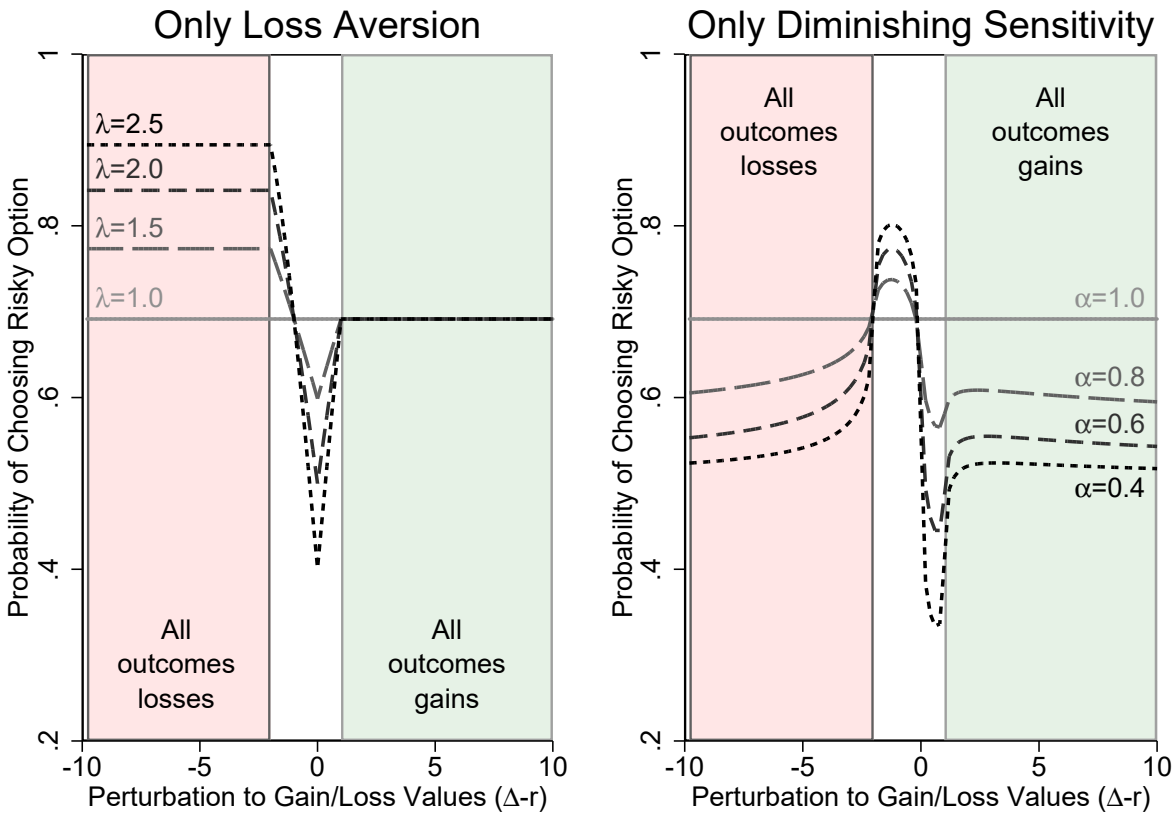
the observed-utility case, this results in a pattern of parallel lines of slope 1.

Despite our desire to not rely on the functional form of the choice probability function in our testing approach, we note that the choice probability function shows distinctive patterns when generated by variants of Prospect Theory. To help explain the pattern seen in the left panel of Figure 2, it is helpful to isolate the role of individual pieces of prospect theory. Figure 3 presents alternative choice-probability functions that are simulated with only loss aversion present or with only diminishing sensitivity present. When assessing these choice-probability functions, it is useful to consider their structure over three different regions: the region in which all gamble outcomes fall in the loss domain, the region in which all gamble outcomes fall in the gain domain, and the region in which the risky gamble involves a potential for gain or loss. These regions are shaded in the panels of Figure 3.

As seen in the left panel of Figure 3, the choice-probability function influenced by loss aversion alone has two distinctive features. First, there is a v-shaped dip in the probability of choosing the risky option in the center of the figure. This dip occurs over the range of $\Delta - r$ values where the reference point falls between the good and bad outcomes of the risky gamble. When this happens, the kink that occurs at the reference point causes first-order risk aversion, leading to the decline it the probability of choosing the risky option. The influence of the kink is small when it occurs near the ends of the range of options considered and is maximized at at an intermediate point, leading to the v shape. Outside of the range of $\Delta - r$ values mapping to this v-shaped dip, either all outcomes of either gamble map to gains or all outcomes of either gamble map to losses. With diminishing sensitivity "turned off," the utility function is linear and thus locally risk neutral in either of these regions. Because the risky option has a positive expected value, it is favored under a risk neutral evaluation, and the stable probability of choosing the risky option reflected within this region is determined by the probability that this utility evaluation is not overturned by unfavorable draws of the random-utility error terms. The higher rate of choosing the risky option over losses arises because the steeper utility function over losses leads to a greater difference in the deterministic utility component, thus leading to a lower probability that the error draws overturn the deterministic evaluation.

As seen in the right panel of Figure 3, the choice-probability function influenced by

Figure 3: Choice Probabilities Generated by Components of Prospect Theory



Notes: This figure presents examples of choice-probability functions that help isolate the implications of individual components of prospect theory. Both examples are simulated according to the procedure used for Figure 2, but varying the terms α or λ from equation 7. The left figure presents the choice-probability function when loss aversion parameter λ is varied and diminishing sensitivity parameter α is set to 1, capturing cases with loss aversion but no diminishing sensitivity. The right figure presents the choice-probability function when loss aversion parameter λ is set to 1 and diminishing sensitivity parameter is varied, capturing cases with diminishing sensitivity but no loss aversion.

diminishing sensitivity alone similarly has distinctive features. First, notice that scaling up the role of diminishing sensitivity reduces the scale of the deterministic utility component, thus leading to a greater role of the random utility term and choice probabilities closer to 50-50. Next, notice that diminishing sensitivity generates risk loving over losses and risk aversion over gains. This generates a greater propensity to choose the risky option in the "all outcomes losses" region than the "all outcomes gains" region. In either region, as the magnitude of gains or losses gets more extreme (i.e., $|\Delta - r|$ gets large), sensitivity diminishes

and the magnitude of the deterministic component of utility differences declines. This leads the random-utility error terms to have a larger role in determining outcomes, pushing choice probabilities towards 50%. In the intermediate region in which the risky option outcomes span the reference point, effective risk preferences transition from risk loving to risk aversion. On either side of this region they are briefly amplified due to the steep region of the utility curve near zero, generating the accentuated sinusoidal pattern.

While the precise shape of the resulting choice function will depend critically on the parametric assumptions that are imposed, the qualitative patterns discussed here may be used as a means of assessing whether choice data are consistent with the predictions of typical parameterizations of Prospect Theory. Returning to our representation of primary interest, we note that the prediction of parallel slope-1 level-sets (reflected in the left panel of Figure 2) was generated with minimal functional form assumptions. Parametric assumptions about diminishing sensitivity or loss aversion merely change the utility values associated with each parallel line, but do not change the parallel-line structure itself.

1.2.2 Predictions when the Reference Point is Misspecified

Proposition 1 indicates that the conditional expectation $E[Y|\Delta, r^c]$ has a distinctive single-index structure when the candidate reference point is correctly specified. We next present two results that illustrate how this structure can be distinguished from what we would see when an incorrect candidate referent is examined, either because another reference point is used or because decisions are not reference dependent.

Proposition 2. ***No Single-Index Representation for Incorrectly Specified Referent:*** *Consider choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. Assume that choices are governed by reference-dependent utility (Assumption 1) with locally linear direct utility (Assumption 2). Let r^c be a candidate reference point that satisfies $r^c \perp \Delta - r$ and $r^c \perp \epsilon$. Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. There does not exist a non-constant function g such that $E[Y|\Delta, r^c] = g(\Delta - r^c)$.*

Proof. Assume for contradiction that there exists a non-constant single-index function g_1 such that $E[Y|\Delta, r^c] = g_1(\Delta - r^c)$. Note that, because $r^c \perp \epsilon$, the line of calculations leading

to the proof of Proposition 1 imply that there exists a function g_2 such that $E[Y|\Delta, r^c, r] = g_2(\Delta - r)$. Applying the law of iterated expectations, we have $g_1(\Delta - r^c) = E[Y|\Delta, r^c] = E[E[Y|\Delta, r^c, r]|\Delta, r^c] = E[g_2(\Delta - r)|\Delta, r^c]$. The assumption that g_1 is nonconstant implies that there exists a triple (Δ_1, r_1^c, r_2^c) such that $g_1(\Delta_1 - r_1^c) \neq g_1(\Delta_1 - r_2^c)$. This implies that $E[g_2(\Delta - r)|\Delta_1, r_1^c] \neq E[g_2(\Delta - r)|\Delta_1, r_2^c]$, which contradicts the assumption that $r^c \perp \Delta - r$. □

Proposition 2 establishes that, in cases where a candidate reference point is simulated to be independent from the gain/loss evaluation, $E[Y|\Delta, r^c]$ will not admit a single-index representation as it does in the case when the homogeneous null hypothesis holds. We have proven this statement under the strong assumption that the false candidate reference point is statistically independent from the our variation in gain/loss amounts $(\Delta - r)$ because it makes the logic of the proof particularly straightforward: when we consider a candidate reference point that is independent of the true gain/loss evaluation, choice probabilities will not react to variation in that potential reference point. In our experiment, we will generate candidate reference points in a manner that satisfies this independence assumption and we will directly assess whether choice probabilities vary in manner consistent with these predictions. While these reasons make imposing statistically independent referents convenient, they are not strictly necessary—the lack of a single-index representation can be proven under significantly weaker requirements, although with more assumed structure necessary to facilitate the proof.[8]

Consider next the case when decisions are not reference dependent.

Assumption 4. ***Non-Reference-Dependent Utility:*** *Consider two gambles, $\mathcal{G}_0$ and $\mathcal{G}_1$. $\mathcal{G}_1$ is chosen over $\mathcal{G}_0$ only if $U(\mathcal{G}_1) \geq U(\mathcal{G}_0)$, where*

$$U(\mathcal{G}) = \sum_{o \in O^{\mathcal{G}}} p_o^{\mathcal{G}} \cdot \psi(c_o^{\mathcal{G}}) + \epsilon. \tag{8}$$

ψ is an increasing function that represents standard Bernoulli utility. ϵ is the realization of an i.i.d random-utility term distributed according to the probability density function f_ϵ.

[8]Note, however, that our approach would not be suitable for distinguishing r_1 and r_2 when $r_1 = r_2 + c$ for some constant c.

Proposition 3. ***Degenerate Single-Index Representation for Non-Reference-Dependent Decisions:*** *Consider choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. Assume that choices are governed by non-reference-dependent utility (Assumption 4) with locally linear direct utility (Assumption 2). Let r^c be a candidate reference point. Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. $E[Y|\Delta, r^c] = \bar{Y}$ for some constant value $\bar{Y}$.*

Proof. Notice that the non-reference-dependent utility function is simply the reference-dependent utility function of Assumption 1 with the constraint that $\phi(x) = 0$ for all x. To assess the predicted structure of $E[Y|\Delta, r^c]$ in a case where the individual is not reference dependent, we may reconsider the calculations presented in equations 4-6 while setting all ϕ terms to zero. If these terms are all zero, the consolidated single-index function ν also is zero. Equation 6 then becomes $E[Y|\Delta, r] = Pr(k \geq \epsilon_0 - \epsilon_1)$, implying that our conditional expectation of interest is constant in Δ and r^c. □

Because constants can be represented as single-index functions (where the function maps any value of the single index to the same constant), an implication of Proposition 3 is that even non-reference-dependent utility results in a single-index representation of $E[Y|\Delta, r^c]$. However, this representation is guaranteed to be a simple constant function, and therefore non-reference-dependent utility would be ruled out in the case where the rationalizing single-index function were non-constant.

1.2.3 Testing in the Presence of Heterogeneity

The results thus far may be interpreted as providing conditions for testing models under homogeneity assumptions. For a given utility function, the statistical object of interest, $E[Y|\Delta, r^c]$, has different predicted features when the candidate reference point the true reference point, when the candidate reference point is not the true reference point, or when utility is not reference dependent. In this section, we consider the implications of heterogeneous utility parameters, heterogeneous choice environments, and heterogeneity in reference-dependent versus non-reference dependent utility adoption. Our approach has desirable robustness to all three types of heterogeneity, derived from the same property of single index

models expressed in Lemma 1.

Lemma 1. *Say $E[Y|\Delta, r^c, \theta] = g_\theta(\Delta - r^c)$ for all $\theta \in \Theta$, where θ indexes a finite set. If the distributions of Δ and r are statistically independent from θ, there exists a function g such that $E[Y|\Delta, r^c] = g(\Delta - r^c)$.*

Proof. Applying the law of iterated expectations, notice that $E[Y|\Delta, r^c] = E[E[Y|\Delta, r^c, \theta]|\Delta, r^c] = E[g_\theta(\Delta - r^c)|\Delta, r^c]$. Using the assumption that θ is statistically independent from Δ and r, we may now define a new single-index function $\bar{g}(\Delta - r^c) = E[g_\theta(\Delta - r^c)|\Delta, r^c] = E[Y|\Delta, r^c]$. □

Put simply, this proposition guarantees that, in a population of heterogeneous types each endowed with a different single-index function, another averaged single-index function will rationalize the population-wide conditional expectation. This has immediate and strong implications for the robustness of our approach to presence of several types of heterogeneity. The following propositions all follow immediately from Lemma 1 after applying previous propositions to establish the existence of single-index representations within groups.

Proposition 4. ***Robustness to Heterogeneity in Utility Parameterizations:*** *Consider choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. Assume that choices between gambles are presented to a heterogeneous population of decision makers indexed by their type $\theta \in \Theta$. All decision-makers satisfy Assumptions 1, 2, and 3, with no further homogeneity assumptions on the components of the utility functions. Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. If the distributions of Δ and r are statistically independent from θ, there exists a function g such that $E[Y|\Delta, r] = g(\Delta - r)$.*

Proposition 4 implies that the approach suggested by Proposition 1 remains valid for a population of agents who are *heterogeneous in parameterizations of reference-dependence*. The key requirement is that agents all satisfy Assumptions 1, 2, and 3, regardless of the presence of heterogeneity in the individual utility components. If individuals differ in the structure of their direct utility functions (ψ), their reference-dependent utility functions (ϕ) or the distribution of their individual utility shocks (f_ϵ), the single-index function guaranteed to exist by Proposition 1 will differ across agents. However, when the conditional expectation of Y is estimated from data that pools these heterogeneous agents, it can be explained by

the single-index function formed by averaging the type-specific functions. Testing for single-index structure continues to provide a means for rejecting the model.

Proposition 5. ***Robustness to Heterogeneity in Base Gambles:*** *Consider a population of decision makers indexed by their type $\theta \in \Theta$, where each type faces choices between gambles $S(\Delta|\mathcal{G}_0^\theta)$ and $S(\Delta|\mathcal{G}_1^\theta)$. All decision-makers satisfy Assumptions 1, 2, and 3. Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1^\theta)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0^\theta)$ is chosen. If the distributions of Δ and r are statistically independent from θ, there exists a function g such that $E[Y|\Delta, r] = g(\Delta - r)$.*

Proposition 5 implies that the approach suggested by Proposition 1 remains valid for a population of agents who are *heterogeneous in their base gambles.* While it is important that each type face Δ-shifted gambles, the base gambles off which Δ-shifting occurs may vary across groups without jeopardizing the single-index-function-based approach to identification.

Proposition 6. ***Robustness to Heterogeneity in Reference Dependence:*** *Consider a population of decision makers of two types, θ^{RD} and θ^{NRD}. Each type faces choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. θ^{RD} decision-makers satisfy Assumptions 1, 2, and 3. θ^{NRD} decision-makers satisfy Assumptions 4 and 2. Let variable Y take the value of 1 if $S(\Delta|\mathcal{G}_1)$ is chosen and the value of 0 if $S(\Delta|\mathcal{G}_0)$ is chosen. If the distributions of Δ and r are statistically independent from θ, there exists a nonconstant function g such that $E[Y|\Delta, r] = g(\Delta - r)$.*

Proposition 6 implies that the approach suggested by Proposition 1 remains valid for a population of agents who are *heterogeneous in the presence of reference dependence.* If the reference point is correctly specified for the individuals who are reference dependent, a non-constant single-index representation of $E[Y|\Delta, r^c]$ will still arise regardless of the presence of non-reference-dependent types.

1.3 Summary

In this section, we have documented a key distinguishing feature of reference-dependent choices: in appropriately structured choice environments, they result in choice probability

functions with distinctive level-sets. These level-sets reflect a simple intuition: that if individuals care about gain/loss evaluations, then raising all consumption amounts by a fixed amount can be offset by raising the reference point by the same amount. This feature arises with only minimal structure placed on the assumed utility model and is robust to a variety of important forms of heterogeneity. While this property can be informally assessed graphically—for example, by generating a contour plot and assessing if it is "close enough" to having the predicted parallel-line structure—it is also amenable to formal statistical testing. In the next section we present our approach to conducting that test.

2 Proposed Testing Strategy

As we established in the previous section, in a broad class of reference-dependent models, conditional choice probabilities admit a non-constant single-index representation if the reference point has been correctly specified. This observation allows us to link the behavioral economic literature on reference dependence to an econometric theory literature on specification testing of single-index models. In this section, we present our formal testing strategy that arises from that linkage.

Our econometric approach builds on Fan and Li's (1996) nonparametric test of single-index specification. Conceptually, when applied to a function of two variables, their test involves estimating a kernel-smoothed approximation of $E[Y|x_1, x_2] = g_1(x_1, x_2)$ that does not impose single-index structure and comparing that to a kernel-smoothed approximation of $E[Y|x_1, x_2] = g_2(\beta_1 x_1 + \beta_2 x_2)$ that does impose the single index structure. To modify this to our setting, we simply take advantage of the additional restrictions imposed by a reference-dependent model (i.e., that $\beta_1 = -\beta_2$). Additionally, we modify the test to allow for clustered observations as opposed to an i.i.d. sample, which is needed to make our test applicable in experiments eliciting multiple evaluations per subject and assigning reference points at the subject level.

In the interest of making our approach accessible to readers with limited training in, or fundamental interest in, nonparametric econometrics, the presentation of our test statistics in this section is extremely concise. Full details of the derivation of our test statistic, complete

with necessary technical assumptions and derivation of the asymptotic distribution, are available in the Appendix A.

2.1 The Stage-1 Test

The goal of this analysis is to test whether $E[Y|\Delta, r]$ admits a single-index representation. We do so by assessing a finite-sample estimate of

$$E[v \cdot f(\Delta - r)E[v \cdot f(\Delta - r)|\Delta, r]] \tag{9}$$

where $f(\Delta - r)$ is the p.d.f. of $(\Delta - r)$ and v is the expected approximation error induced by assuming a single-index representation (i.e., $v = E[Y|\Delta, r] - E[Y|\Delta - r]$). Note that if $E[Y|\Delta, r]$ admits a single index representation, v is zero for any Δ and r. Consequently, the full expectation evaluated in equation 9 will also be zero. Given finite-sample kernel-based approximations to both $E[Y|\Delta, r]$ and $E[Y|\Delta - r]$, approximation error will cause v to not be identically zero, but instead distributed around zero, as too will be the full expectation evaluated in equation 9. Our test proceeds by generating a test statistic that, with proper scaling, is asymptotically normally distributed under the null hypothesis of a admission of a single-index representation but that diverges under the alternative hypothesis.

The key statistic that estimates equation 9 is:

$$\Omega = \sum_i \sum_j \frac{1}{N}[\bar{v}_{(i,j)}\widehat{f}_a(\Delta_{(i,j)} - r_i)] \sum_{i' \neq i} \sum_{j'} \frac{1}{(N-m)}[\bar{v}_{(i',j')}\widehat{f}_a(\Delta_{(i',j')} - r_{i'})][\frac{1}{h^2}K^h_{(i,j),(i',j')}]. \tag{10}$$

In this equation, i indexes the subject of interest and i' indexes other subjects. j is used to index the choice of subject i, and j' is used to index the choice of subject i'. n denotes the number of subjects (i.e., clusters), and m denotes the number of observations per subject, yielding total sample size of $N = n \cdot m$. In this equation, $\bar{v}_{(i,j)} \equiv Y_{(i,j)} - \widehat{E}[Y_{(i,j)}|\Delta_{(i,j)} - r_i]$ is a kernel-based estimate of the approximation error from a best-fit single-index model. $\widehat{f}_a(\Delta_{(i,j)} - r_i)$ is a kernel-based estimate of $f(\Delta - r)$ with bandwidth a. $K^h_{(i,j),(i',j')} \equiv k(\frac{\Delta_{(i,j)} - \Delta_{(i',j')}}{h})k(\frac{r_i - r_j}{h})$ is a kernel-based estimate of the the joint distribution of Δ and r with bandwidth h, formed

as a product of two univariate Gaussian kernels $k(.)$.[9]

In Equation 10, the inner summations (i.e., the component beginning with $\sum_{i' \neq i} \sum_{j'}$) provide a empirical analog of the conditional weighted approximation error ($E[vf(\Delta - r)|\Delta, r]$) in Equation 9. This term is multiplied by $\bar{v}_{(i,j)} \widehat{f}_a(\Delta_{(i,j)} - r_i)$, which serves as the empirical analog to the $v \cdot f(\Delta - r)$ term in Equation 9. This product is averaged over all choices j made by each subject i, providing an estimate of the outer unconditioned expectation operation in equation 9.

In Appendix A, we prove that under the null hypothesis that $E[Y|\Delta, r]$ admits a single index representation (and while imposing standard regularity conditions for kernel estimation), $Nh\Omega$ is asymptotically normally distributed with a mean of zero and a standard deviation consistently estimated by $\sqrt{2 \cdot (\widehat{\sigma_a^2} + \widehat{\rho_a^2})}$, where

$$\widehat{\sigma_a^2} = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} [\bar{v}_{(i,j)} \widehat{f}_a(\Delta_{(i,j)} - r_i)]^2 [\bar{v}_{(i',j')} \widehat{f}_a(\Delta_{(i',j')} - r_{i'})]^2 K^h_{(i,j),(i',j')} [\int k(u)^2 du]^2$$

and

$$\widehat{\rho_a^2} = \frac{(m^2-1)h}{N(N-m)(m-1)^2 h^3} \sum_i \sum_{j_1 \neq j_2} \bar{v}_{(i,j_1)} \widehat{f}_a(\Delta_{(i,j_1)} - r_i) \bar{v}_{(i,j_2)} \widehat{f}_a(\Delta_{i,j_2} - r_i) \sum_{i' \neq i} \sum_{j_1' \neq j_2'}$$
$$\bar{v}_{(i',j_1')} \widehat{f}_a(\Delta_{(i',j_1')} - r_i') \bar{v}_{(i',j_2')} \widehat{f}_a(\Delta_{(i',j_2')} - r_i') k(\frac{\Delta_{(i',j_1')} - \Delta_{(di,j_1)}}{h}) k(\frac{\Delta_{(i',j_2')} - \Delta_{(i,j_2)}}{h}) k(\frac{r_i' - r_i}{h}) \int k^2(u) du.$$

Comparing the value of $\frac{Nh\Omega}{\sqrt{2 \cdot (\widehat{\sigma_a^2} + \widehat{\rho_a^2})}}$ against the standard normal distribution therefore provides the p-values of our test.

Note that, if $\widehat{\rho_a^2}$ is set to zero, this result follows closely from Fan and Li (1996) who assume an i.i.d. data generating process. For experimental applications that elicit multiple observations per subject, the i.i.d. assumption can be violated due to the correlations that arise within-subject, and this violation changes the asymptotic distribution. One may interpret the $\widehat{\rho_a^2}$ term as a correction to the original Fan and Li (1996) estimate of variance that corrects for an assumed absence of correlation within subject.

[9] All kernel-based estimates in this paragraph are fully defined in Appendix A, equations 14, 15, and 17.

2.2 The Stage-2 Test

While the stage-1 test directly assesses whether we can statistically reject a single-index representation, it does not impose any constraint that this representation must be non-constant. In the stage-2 test, we assess whether we can statistically reject a representation of $E[Y|\Delta, r]$ as a constant function.

Conceptually, the stage-2 test uses the same approach as the stage-1 test. Consider the approximation error induced by fitting a constant function, denoted $u = E[Y|\Delta - r] - E[Y]$. Our goal is to estimate a quantity similar to equation 9 in Stage 1:

$$E[uf(\Delta - r)E[u|\Delta - r]]. \tag{11}$$

Applying similar approximation methods as in the Stage-1 test, we may generate a finite-sample estimate of equation 11 by

$$\Pi = \sum_i \sum_j \frac{1}{N}(Y_{(i,j)} - \hat{\mu}) \sum_{i' \neq i} \sum_{j'} \frac{1}{(N-m)}(Y_{(i',j')} - \hat{\mu})[\frac{1}{a}k(\frac{X_{(i',j')} - X_{(i,j)}}{a})]. \tag{12}$$

In this equation, $\hat{\mu} = \frac{1}{N}\sum_i \sum_j Y_{(i,j)}$, $X_{(i,j)} = \Delta_{(i,j)} - r_i$, $k(.)$ is the univariate Gaussian kernel, and a is the bandwidth.

In Appendix A, we prove that under the null hypothesis that $E[Y|\Delta, r]$ is constant (and while imposing standard regularity conditions for kernel estimation), $N\sqrt{a}\Pi$ is asymptotically normally distributed with a mean of zero and a standard deviation consistently estimated by $\sqrt{2(\hat{\sigma}_\mu^2 + (m^2-1)a\hat{\rho}_\mu^2)}$, where[10]

$$\hat{\sigma}_\mu^2 = \frac{1}{N(N-m)a} \sum_i \sum_j \hat{u}_{(i,j)}^2 \sum_{i' \neq i} \sum_{j'} \hat{u}_{(i',j')}^2 k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \int k^2(u)du$$

[10] As defined in Equation 30, $\hat{u}_{(i,j)}$ is the estimated approximation error for observation (i, j).

and

$$\hat{\rho}_{\mu}^{2} = \frac{1}{N(N-m)(m-1)^{2}a^{2}} \sum_{i} \sum_{j_1 \neq j_2} \hat{u}_{(i,j_1)} \hat{u}_{(i,j_2)} \sum_{i' \neq i} \sum_{j_1' \neq j_2'} \hat{u}_{(i',j_1')} \hat{u}_{(i',j_2')} \\ k\left(\frac{X_{(i',j_1')} - X_{(i',j_2')}}{a}\right) k\left(\frac{X_{(i,j_1)} - X_{(i,j_2)}}{a}\right).$$

Thus, similar to the stage-1 test statistic, comparing the value of $\frac{N\sqrt{a}\Pi}{\sqrt{2(\hat{\sigma}_{\mu}^{2} + (m^{2}-1)a\hat{\rho}_{\mu}^{2})}}$ against the standard normal distribution provides the p-values of our stage-2 test.

2.3 Proposed Usage and Interpretation of the Test

With our two test statistics formally defined, we now describe how these tests could be applied and interpreted.

Consider a situation in which an experimenter has presented subjects with choices between Δ-shifted gambles in the presence of a randomly varied candidate reference point r^{c}. If this experimenter is comfortable with the two fundamental assumptions imposed in Proposition 1—the structure of the random utility model in Assumption 1 and the local-linearity restriction in Assumption 2—then he may attempt to reject the proposed reference point by attempting to statistically reject the non-constant single-index representation that this proposition guarantees. To do so, the experimenter could conduct the stage-1 and stage-2 tests described in this section.

If the stage-1 test rejects the null of a single-index representation, or if the stage-1 test fails to reject the null of a single-index representation but the stage-2 test subsequently fails to reject a representation by a constant function, we interpret our test to have failed. Formally, this means that these two tests taken together do not provide support for the existence of the non-constant single-index representation that Proposition 1 guarantees. If a non-constant single-index representation does not exist, at least one of the assumptions of Proposition 1 does not hold, and if Assumptions 1 and 2 are already conceded then we may infer that the assumption to be rejected is Assumption 3: the null hypothesis of correct reference point specification.

If the stage-1 test fails to reject the null and the stage-2 test rejects the null, we interpret

our test to have passed. Taken together, these two tests imply that the choice probability function is non-constant, but is statistically indistinguishable from a non-constant single-index function. Of course, a failure to reject a null hypothesis does not mean that the null hypothesis is true. However, the parallel level-set patterns that we have documented are rather distinctive. If the researcher assesses there to be strong evidence in support of these patterns, we view that as strong evidence in support of the claim that the tested model of the reference point could indeed be the true one.

Applying this template requires specifying a threshold for rejecting or failing to reject a null hypothesis. Throughout our analysis we adopt the standard 5% α-level.

3 Applying Our Approach in an Experiment

In this section, we demonstrate how to run an experiment optimized for our econometric approach. We begin by discussing experimental design considerations in the abstract, and then present a concrete example of an experiment optimized for this approach.

3.1 Considerations for Experimental Design

The conceptual approach we detailed in Section 1 suggests a straightforward procedure for testing a candidate model of reference points. Roughly speaking, the procedure involves presenting subjects with Δ-shifted gambles while exogenously varying the candidate reference point, then assessing whether the necessary single-index structure is ruled out by the statistical approach detailed in Section 2. In this section, we discuss some key considerations that arise when designing an experiment that executes that strategy.

3.1.1 Designing the Gambles

Utilizing our testing approach requires presenting subjects with Δ-shifted gambles, denoted $S(\Delta|\mathcal{G})$. Given this requirement, the gambles ultimately presented to subjects are fully determined by the experimenter's choice of base gambles ($\mathcal{G}$) and by the procedure for choosing shifts to those gambles (i.e., choosing the distribution of Δ). While we formulated our approach to be applicable with arbitrary choices of these primitives, some versions of

them are preferable to others for creating a desirable experience for subjects or for maximizing statistical power. When making these design decisions, we recommend that experimenters keep two considerations in mind.

First, the potential for a reference point to survive our testing procedure depends critically on presenting gambles in which choice probabilities vary with Δ and r. To illustrate with an extreme example, imagine we presented subjects choices between $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$ where $\mathcal{G}_0$ offers a 100% chance of earning \$1, $\mathcal{G}_1$ offers a 50-50 chance of earning -\$1,000 or -\$2,000, and Δ is drawn from a uniform distribution ranging from -\$1 to +\$1. This choice of gamble structure is not expected to generate informative choice data—regardless of the realization of Δ or r, $S(\Delta|\mathcal{G}_0)$ is so clearly preferable to $S(\Delta|\mathcal{G}_1)$ that we'd expect it to essentially always be chosen, and thus $E[Y|\Delta, r]$ is expected to be approximately constant. In contrast, the most useful choices to present to subjects are those where, for different ranges of Δ within the support of f_Δ, substantially different choice probabilities arise. To illustrate with an example, in the simulation presented in Figure 2 we documented that the choice probability function would significantly vary when Δ-r varied between -5 and +5. When considering potential gambles to present, we recommend conducting similar such simulations, determining the region of Δ-r values over which variation in choice probabilities occurs for standard utility parametrizations, and then choosing the sampling distribution of Δ such that the realizations of Δ-r will fall in that range.

Second, from a subject-engagement perspective, it may be desirable to obfuscate the structured manner in which gambles are being randomly generated. Concretely, consider a situation where a subject is presented with 10 choices between gambles, all generated from the same fixed base gambles. As in the example presented in Figure 2, imagine the base gambles present a "safe option" offering \$0 with certainty and "risky option" offering a 50-50 chance of +\$2 or -\$1. The sequence of choices presented to the subject could be:

Decision 1: +\$2 with certainty vs. a 50-50 chance of +\$4 or +\$1 (i.e., $\Delta = 2$),

Decision 2: -\$1.5 with certainty vs. a 50-50 chance of +\$0.5 or -\$2.5 (i.e., $\Delta = -1.5$),

Decision 3: +\$0.6 with certainty vs. a 50-50 chance of +\$2.6 or -\$0.4 (i.e., $\Delta = 0.6$),

Decision 4: -\$2 with certainty vs. a 50-50 chance of +\$0 or -\$3 (i.e., $\Delta = -2$),

... and so on.

When faced with a sequence like this, we believe that many attentive subjects will relatively quickly notice that all decisions follow the same structure, differing only by all consumption amounts shifting up and down in unison. We worry that this realization could invite subjects to disengage from the experiment due to boredom with subtle variants of the same question. One means of dealing with this issue, which we adopt in our experiment and we recommend, is to present choices based on several different base gambles.[11] If the order of presentation is alternated or randomized, this makes quickly inferring the pattern substantially more difficult for subjects. In practice, we believe this can be done in a manner that results in all gambles seeming fully randomly generated from the perspective of subjects, which is desirable for motivating consistent attention to each new choice presented. Using a set of different base gambles also allows the researcher to present gambles that are predicted to be close to marginal for a range of different assumptions on utility structures.

3.1.2 Varying Reference Points

Utilizing our testing approach additionally requires generating exogenous variation in the candidate reference point that one wishes to test. When designing an experiment that uses this approach, the experimenter must decide how to induce this variation and what distribution of variation to induce.

The decision of how to induce variation in the reference point is fundamentally tied to the specific model of the referent being tested, and will vary across models. For at least some reference points, saliently presenting the reference value may be enough to set it according to an experimenter's goals. For example, in our experiment, we test whether subjects adopt externally suggested goals or reported group average earnings as reference points. Prior literature (referenced in the introduction) suggests that salient presentations of these objects could be sufficient as an experimental manipulation. For other models of reference point formation, more laborious designs may be needed. For example, when testing an endowment-based reference point, the experiment may require randomizing the physical

[11] Recall that Proposition 5 established that pooling choices derived from several different base gambles is viable with our testing strategy.

provision of a good. In the interest of providing a technique that applies to general models of reference-point formation, we do not take a stand on the randomization technique that is required. Instead, we merely emphasize that the approach presented in this paper requires that r and Δ be capable of varying independently from one another. This is possible for nearly all candidate reference points discussed in the literature, but it is notably not possible in the model of Kőszegi & Rabin (2006). Note, however, that a test of Kőszegi-Rabin reference points can be generated in a simple extension of our framework. We provide this approach in Appendix B.[12]

Upon determining a means of randomizing the referent, the researcher must then decide whether to randomize reference points within subject or between subjects. There are two natural paradigms to follow, with either being valid from the perspective of our statistical approach. One option is to assign Δ and r are randomly for each individual decision made. Another option is to assign Δ randomly for each individual decision, but to assign a single, fixed value of r to each experimental subject. We follow the later option in our experiment, and believe it will typically be preferable. While random assignment of reference points at the decision level has advantages for statistical power, we worry that varying a reference point across sequential decisions would feel very unnatural for subjects and might make them attend to and rely on the referent less than they would if it were stable.

When determining the distribution of variation in the referent to induce, the experiment faces the same considerations that were just presented for gamble choice. Conditional on testing the true reference point, establishing the presence of a non-constant single-index representation is easiest when $\Delta - r$ is sampled from a region with substantial variation in $E[Y|\Delta, r]$. We recommend that the choice of the sampling distribution of the reference point be made in conjunction with the choice of base gambles and the distribution of Δ to achieve that goal, following the guidance provided in Section 3.1.1.

[12]Due to its popularity with behavioral economic theorists, some perceive the Kőszegi-Rabin model to be an especially important candidate for testing. We note, however, that it is adopted in a small minority of empirical tests: the recent metaanalysis of Brown et al. (2020) examines 522 empirical estimates of loss aversion and finds that only 18 of them applied expectations-based reference points. Furthermore, this model now has well documented limitations in explaining experimental data (see, e.g., Heffetz & List, 2014). While we do view the Kőszegi-Rabin notion of reference points as a reasonable candidate for testing, we do not view it as elevated above other candidate theories.

3.1.3 Assessing Power

In any experiment, it is crucial to generate a sufficient sample size for powered statistical analysis. In our experiment, this means ensuring that the two-stage testing procedure detailed in Section 2 rejects false candidate reference points with high probability and rejects true candidate reference points with low probability. To guide our experimental design, we assessed these issues in a large-scale simulation study. This study involved simulating choices among the gambles that we present for a wide range of possible utility parameterizations. Based on these analyses, we determined that collecting 4 observations per subject for 300 subjects resulted in tolerable rates of type-1 and type-2 error.[13] At this sample size, across the different parameterizations we considered, our estimated rate of "passing" the true reference point was 95.5% on median (and above 75.0% for 90% of parameter combinations). Our estimated rate of "passing" the false reference point was 4.0% on median, and below 8.5% in 90% of parameter combinations. We provide full details of our simulation study in Appendix C, and we recommend researchers follow the template of this analysis prior to running their own experiments.

3.2 Design of Our Experiment

In our experiment, subjects were presented with a series of choices between a sure option and a risky option. Each option was presented as a gamble based on the flip of a fair coin. For the risky option, heads and tails mapped to different amounts of money, whereas for the sure option heads and tails mapped to the same amount of money.

After the initial presentation of the format of decisions, subjects were told that they would face 20 decisions of this type. They were also told that one of these decisions would be randomly selected to be the decision that "counts." For that decision, we would simulate a coinflip and deliver a bonus payment as dictated in the the gamble that they chose. Payment for taking the full study consisted of a $4 fixed payment plus this bonus. These instructions were followed by a series of three questions meant to verify their understanding of the decision

[13]Note that our experiment presents subjects with 20 choices: 4 Δ-shifts of each of 5 different base scenarios. While we can conduct our test pooling all choices together, we wished to ensure that we would be powered to run our test on the unpooled data from each of the 5 different base scenarios.

format and correct any misunderstanding that still existed. Subjects were presented with an example gamble followed by two questions asking them to verify the amounts of money they could earn if they selected option A or option B. They faced a third multiple-choice question that asked them to indicate the manner in which they would be compensated for the study to ensure they understood the random selection of a decision that "counts." After answering these questions subjects were given feedback on their responses and told the correct answer if they answered incorrectly.

A final introductory screen introduced potential reference points to subjects. Subjects were told:

Starting on the next screen, you will face the series of choices that were just described. To decide which option to choose, participants sometimes find it useful to use benchmarks for their earnings.

- *Some participants find it helpful to set goals for themselves when completing these tasks. We would like for you to view earning at least a $[R1] bonus as your goal.*
- *Some participants find it helpful to compare their performance against averages. We would like for you to imagine that you are part of a group of participants who earned an average bonus of $[R2].*

In this text, the terms $[R1]$ and $[R2]$ are placeholders for the randomly generated values of the potential reference points. Each reference point is populated by i.i.d. draws from a random normal distribution with a mean of \$3.4 (the average value of the sure payment) and a standard deviation of \$0.7.

This screen provides the first mention of the two reference points considered in this study. In all conditions, these two reference points are randomly generated and presented on this page. After this page, subjects move to making their 20 gamble choices under one of three different conditions. One serves as a control condition, in which these reference points are not mentioned again throughout the study. The other two conditions correspond to cases where one of the two reference points is made salient during gamble choices. As illustrated in Figure 4, this salience is achieved by including large red text over the choice interface reminding the subject of either their goal or the average earnings. In those conditions, one

Figure 4: Screenshot of Explanation of Gamble Interface

Notes: This figure presents a screenshot of a gamble choice in our experiment. In this treatment arm, average earnings were made salient as a potential referent. In other treatment arms, the box containing the red text could instead report the goal assigned to the subject, or it could be omitted entirely.

potential reference point is entirely ignored after its first mention in the example text above, whereas the other potential referent is consistently and vividly present throughout the study.

The 20 decisions presented to the subjects differ only in the amounts of money corresponding to the sure payment and the heads and tails outcomes of the risky payment. These amounts were generated in five question groups of four questions each. Within each question group, all gambles are Δ-shifted from a common base gamble but with different i.i.d. draws of Δ from a normal distribution with a mean of zero and a standard deviation of \$0.25. The base gambles used vary by question group and are presented in Table 1.

After subjects made their sequence of 20 gamble choices, they were shown the choice that was randomly selected for incentivization. The gamble was simulated and the subject was informed of their earnings in the study.

Complete text of the experiment, along with details of all data collected, are available in the UAS Experimental Codebook.[14]

[14] Available at https://uasdata.usc.edu/survey/UAS+287.

Table 1: Baseline Gambles

	Sure Amount	*50-50 Values*	
Base Scenario	q_a	q_b	q_c
1	\$3.4	\$2.00	\$4.80
2	\$3.4	\$2.25	\$4.65
3	\$3.4	\$2.45	\$4.65
4	\$3.4	\$2.30	\$4.90
5	\$3.4	\$2.50	\$4.50

Notes: This table presents the payoff values for the five pairs of base gambles considered in experiment.

3.3 Experimental Deployment

In December 2020 and January 2021, we deployed our experiment in the Understanding America Study (UAS), an online panel of American Households.[15] To achieve our targeted sample size of 1,000 responses, the UAS drew a random subsample of 1,333 respondents from their full panel. These 1,333 respondents received invitations to take our study. The study was closed shortly after the target sample size was attained, ultimately resulting in 1,001 complete observations and a 75% response rate.

Panelists in the UAS are recruited in a manner meant to generate samples representative of the United States population. As a result, our sample appears reasonably representative based on observables, although some notable difference exist. Relative to the full U.S. population, participants in our survey are more likely to be female, married, white, and highly educated. Comparing the demographics of those who completed our survey versus those who were invited but did not complete it, we see some evidence of selection for respondents who are Hispanic or Latino, older, and married. (Table D.1 provides full demographic information on our sample.)

Prior to deployment, our study was preregistered on aspredicted.org.[16] This preregistration specified our sample size, precise analyses of interest, and default values for the tuning parameters in our non-parametric approach.[17]

[15] For a detailed description of the UAS, see Alattar et al. (2018).

[16] Available at https://aspredicted.org/7pc6i.pdf.

[17] As specified in the pre-registration, the bandwidth when estimating the two-dimensional density of (Δ, r) in the stage-1 test is $1.5\hat{\sigma}N^{-0.45}$. The bandwidth when estimating the one-dimensional density of $(\Delta - r)$ in the stage-1 test is $1.5\hat{\sigma}N^{-0.35}$. The bandwidth for the one-dimensional density in the stage-2 test is

Table 2: Test Results

Reference Point Tested	Reference Point Made Salient in Treatment Arm		
	None	**Goal**	**Avg Earnings**
Goal	S1: p=0.26 S2: p=0.36	S1: p=0.78 S2: p=0.00	S1: p=0.49 S2: p=0.69
Avg Earnings	S1: p=0.11 S2: p=0.46	S1: p=0.51 S2: p=0.52	S1: p=0.70 S2: p=0.16

Notes: We test whether each reference point presented in our experiment (indicated in the left column) can rationalize the choices made in each treatment arm of the experimented (indicated in the header). In each cell we report the p-values of both the stage 1 (S1) and stage 2 (S2) test statistics, and we color coded the cell to indicate whether the test as a whole passed (in green) or not (in yellow).

3.4 Experimental Results

We begin our empirical analysis by applying the testing approach detailed in Section 2. Recall that our experiment has three treatment arms: one in which the randomly generated goal is made salient in all decisions, one in which the randomly generated average earnings is made salient in all decisions, and a control treatment in which neither potential referent is made salient after its brief initial presentation. Using this structure, we may apply our approach to test if either reference point is adopted in each treatment arm. Because there are multiple applications of either reference point in the existing literature, we believe that either could be adopted as a reference point under the right conditions. We expected that their presentation in the salient condition could be sufficient. However, we also expected that each candidate would not be adopted when it was not made salient.

Table 2 presents the results of our test when applied to the different reference points (indicated in the left column) using the data from different treatment arms (indicated in the header). Cells of this table contain the p-values of the stage 1 (S1) and stage 2 (S2) test statistics, and are color-coded to indicate whether the test passed (in green) or failed (in

$1.5\hat{\sigma}N^{-0.5}$. $\hat{\sigma}$ is the estimated standard deviation of corresponding variables (Δ and r in two-dimensional smoothing in stage 1, $\Delta - r$ for one-dimensional smoothing in stage 1 and stage 2). These bandwidths where choosen with several considerations in mind. First, they were selected to conform with Assumption 7 in Appendix A. Second, they were informed by Fan and Li's discussion of the value of undersmoothing the two-dimensional alternative relative to the one-dimensional null. Finally, the scaling parameter, 1.5, was chosen based on simulated test performance, and is comparable to the scale parameter in other related papers (see Henderson & Parmeter (2015) for a review).

yellow).[18]

Focusing first on the first column, we see that neither reference point passed our test in the control arm in which neither was salient. This conforms with our expectations, and indicates that when a goal or an average earnings value is randomly generated and only very subtlety presented (with no reminders after a single appearance on an instruction screen), subjects do not adopt it as their reference point.

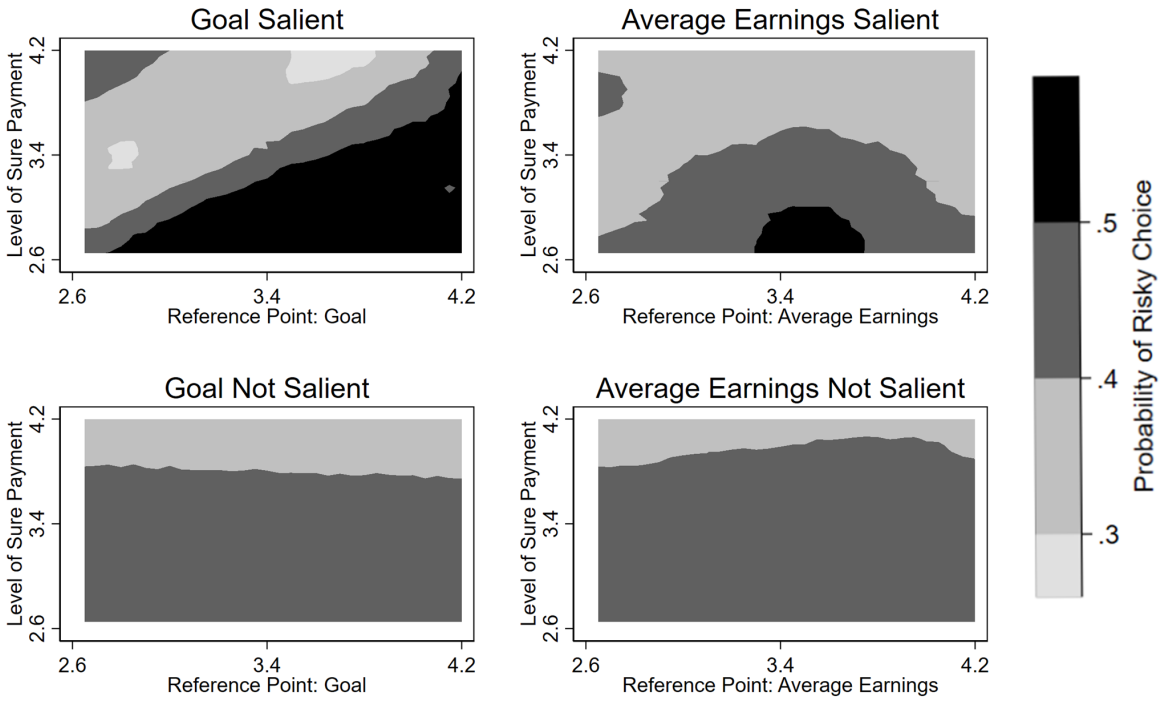
The next column presents results from the treatment arm where goals are made salient. When goals are made salient, goals pass our test as a candidate reference point and average earnings do not. This too conforms with our expectations: existing literature suggests that goals can serve as reference points in at least some circumstances, and we designed this experiment expecting the salient presentation in our design to be sufficient to lead them to be adopted. If goals are adopted as the reference point, this would imply that average earnings are not, and accordingly the average earnings reference point fails to pass the test in this treatment arm.

The final column presents results from the treatment arm where average earnings are made salient. When average earnings are made salient, neither reference point passes our test. This finding may be considered surprising for readers familiar with existing literature on average-based reference points. In the next section, we discuss the implications of this finding for that literature and our advised use of this finding going forward.

The econometric approach used to generate these test statistics applies econometric techniques that are relatively complex and unfamiliar to many readers in the experimental and behavioral literatures. This invites the criticism that it operates as a "black box." However, as we have documented in this paper, these tests may be understood to assess a simple and easy-to-visualize structure on the level-sets of the choice probability function. To illustrate this connection, we next directly examine contour plots of choice probabilities mapped over $\Delta \times r$ space to search for the parallel-line, slope-1 level sets that were the key identifying feature highlighted in our intuitive description of our approach.

[18] This table presents results of our test using pooled data from all 5 base scenarios presented to subjects. In Appendix Table D.2, we present results from these tests when conducted within each of the 5 base scenarios. Across these 30 tests (2 reference points x 3 treatment arms x 5 base scenarios), the base-scenario-specific analyses lead to the same pass/fail conclusion as the pooled test in all but two cases.

Figure 5: Level Sets of Choice Probabilities



Notes: This figure presents contour plots of the conditional probability of choosing the risky option as a function of the level of sure payment (which is always $\$3.4 + \Delta$) and different candidate reference points. The plots on the left apply the goal value as the candidate reference point and the plots on the right apply the average value as the candidate reference point. In the top row, the data are restricted to the treatment arm where the candidate reference point was made salient. In the bottom row, the data are restricted to the two treatment arms where the candidate reference point was not made salient. In all figures, values are derived by local-linear kernel regression of a dummy variable indicating choosing the risky gamble on the variables plotted on each axis. Kernel: Epanechnikov. Bandwidth values are chosen to minimize the integrated mean squared error of the prediction.

To non-parametrically assess the shape of level sets of our choice probability functions, we conduct local-linear kernel regressions of a dummy variable indicating choosing the risky gamble on Δ and r. Figure 5 presents these estimated choice probabilities plotted over a fine grid. For each of our two candidate reference points, we separately conduct this exercise for the treatment arm where the relevant reference point was salient and pooling the two other treatment arms when the relevant reference point was not salient.

We first direct attention to the top left panel, which plots variation in choice probabilities over our randomly generated goal reference points. Despite the completely non-parametric

manner in which this figure has been generated, several clear parallel lines of slope close to 1 are readily apparent. The overall structure of this figure bears remarkable similarity to the example plotted in Figure 2, and serves as a clear demonstration of the patterns we have isolated as hallmarks of a correctly specified reference point. The fact the salient goal reference point passes our test can be understood to derive directly from this pattern: the empirical relationship is "close enough" to the theoretical prediction under a correctly specified reference point that correct specification cannot be rejected.

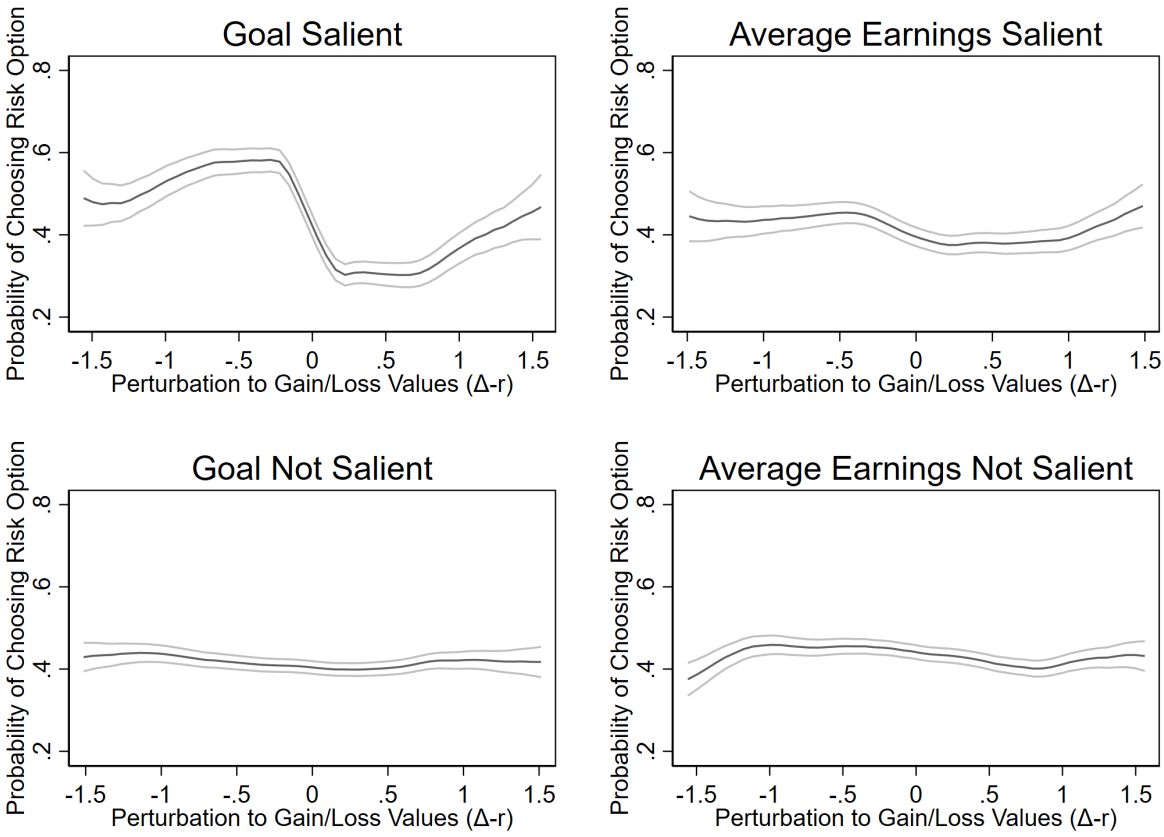
Next turn attention to the bottom left panel, which considers the relationship between choice probability and the goal reference point in the treatment arms where goals are not salient. In contrast to the previously considered panel, this figure does not feature slope-1 parallel lines that were illustrated in Figure 2. When goals are not made salient, the absence of this feature in our data contributes to the rejection of the reference point in our formal statistical tests.

The two panels on the right of this figure present results when average earnings are used as the candidate reference point. As in the panel analyzing non-salient goals, the analysis in these two panels bears little resemblance to the predicted structure under a correctly specified reference point. Again, we see no suggestion of parallel lines of slope 1, and the empirical patterns are "different enough" that they drive the rejection of this reference point in our formal statistical tests.

In the course of developing our test, we have emphasized our desire to not use the shape of the choice probability function $E[Y|\Delta - r]$ itself as a source of identifying power. Avoiding this in our test enables us to avoid reliance on functional form assumptions, and helps generate our test's robustness to heterogeneity. Despite this desire, we still wished to assess whether the choice probability functions that arose in our setting would indeed confirm to the predictions of standard parameterizations of Prospect Theory.

Figure 6 presents our estimated choice probability functions, generated with a local-linear kernel regression of Y on $\Delta - r$. The top left panel presents the estimated choice probability function in the "goals salient" treatment, assessed using goals as the reference point. Recall that this was the sole condition under which a candidate reference point passed our test. In this case, the estimated choice probability function is clearly non-constant, and conforms to

Figure 6: Estimated Choice Probability Functions



Notes: This figure presents plots of the conditional probability of choosing the risky option as a function of the perturbation to gain/loss values ($\Delta - r$), along with estimated 95% confidence intervals. The plots on the left apply the goal value as the candidate reference point and the plots on the right apply the average value as the candidate reference point. In the top row, the data are restricted to the treatment arm where the candidate reference point was made salient. In the bottom row, the data are restricted to the two treatment arms where the candidate reference point was not made salient. In all figures, the conditional choice probabilities are estimated by local-linear kernel regression of a dummy variable indicating choosing the risky gamble on the variables plotted on each axis. Kernel: Epanechnikov. Bandwidth values are chosen to minimize the integrated mean squared error of the prediction.

the general shape that we documented in the simulated example of Figures 2. In contrast, in all other panels the estimated choice probability function is near constant, and does not bear such striking similiarity to Figure 2. This further helps explain the rejection of our test in these cases, since passing our test requires rationalization by a non-constant single-index choice probability function.[19]

[19]Put differently, this provides a visual means of assessing the failure of the stage-2 test to reject the null

In summary, our test suggests that salient goals can serve as reference points, but rejects the adoption of non-salient goals or average earnings (regardless of salience) in our environment. While these claims are based on our novel econometric test, the passage or failure of our test can be quickly visually assessed in simple plots of choice probability functions and their level sets.

4 Discussion

Reference dependence is among the most well-studied phenomena in behavioral economics. And yet, a complete account of how reference points come to be adopted remains elusive. This paper presents a tool for making progress in this domain. In closing, we provide guidance on a final stage of using this tool: deciding how to proceed after individual experiments run with this framework.

As we emphasized in the introduction, our goal in this project was to design a testing framework that was amenable to iterative use across different contexts. With such a tool in hand, researchers can follow a simplified scientific process, cycling between developing new theories, testing them in a standardized manner, and then using the results to guide the development of future theories to test. Given the results of our first use of this approach, it is now appropriate to take stock and consider the implied theory refinements for future tests.

We view our results as clear confirmation of two natural elements of a general theory of the reference point: that goals can serve as reference points, and that salience can moderate reference point adoption. These ideas are not novel. A number of existing papers consider the role of goals as reference points,[20] and we believe that many experimenters already consider the salience of a potential reference point at the design stage of Prospect Theory experiments (even if they relatively rarely discuss this in papers). Despite this lack of novelty, we believe that there is value in our demonstration of these points with novel methods. And additionally, our ability to recover expected results with our novel method provides some

hypothesis of a constant in these cases.

[20] For references, see footnote 6.

reassurance that our method works as intended.

We also view our results as providing some indication of an area where our theories of reference points need further development. Viewing our test in isolation, one possibility is that beliefs about average earnings never serve as reference points. Moving beyond the consideration of our experiment in isolation, however, we would be quite surprised if this interpretation turned out to be true. We view experimental results such as those of Abeler et al. (2011), Gill & Prowse (2012), and Marzilli Ericson & Fuster (2011) to compellingly suggest at least some environments where these reference points could be active. At the same time, we view experimental results as Heffetz & List (2014) and Heffetz (2021) as suggesting that these reference points fail in at least some other environments. Given this literature, we advise against taking our findings as a systematic rejection of average- or expectation-based reference points, and instead interpret them as a clear demonstration of a case where they failed. What features of our decision environment may explain this failure? One possibility is that the reason averages are adopted as reference points is because they are at times endogenously adopted as goals, but when they are placed side-by-side with an explicit goal as in our experioment that endogeneous process is disrupted. Another possibility is that our manipulation of averages—involving asking subjects to imagine themselves as part of a group with a particular earnings—was insufficient as compared to a situation where groups averages were truly different. These are merely two focal possibility, and other candidate explanations will surely be constructed. Moving forward, we advocate for the formal testing of hypotheses like these, as well as hypotheses regarding the many other candidate reference points that were not included in our experiment. The approach we provide in this paper can be put to good use in that pursuit.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2), 470–92.

Alattar, L., Messel, M., & Rogofsky, D. (2018). An introduction to the Understanding America Study internet panel. *Social Security Bulletin*, 78(2).

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6), 1657–1672.

Barseghyan, L., Molinari, F., O'Donoghue, T., & Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6), 2499–2529.

Brown, A. L., Imai, T., Vieider, F., & Camerer, C. F. (2020). Meta-analysis of empirical estimates of loss-aversion. *CESifo Working Paper No. 8848*.

Chapman, J., Snowberg, E., Wang, S., & Camerer, C. (2018). *Loss Attitudes in the U.S. Population: Evidence from Dynamically Optimized Sequential Experimentation (DOSE)*. Working Paper 25072, National Bureau of Economic Research.

Crawford, V. P. & Meng, J. (2011). New york city cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, 101(5), 1912–32.

DellaVigna, S. (2018). Structural behavioral economics. *Handbook of Behavioral Economics: Applications and Foundations*, 1, 613–723.

Fan, Y. & Li, Q. (1996). Consistent model specification tests: Omitted variables and semi-parametric functional forms. *Econometrica*, 64(4), 865–890.

Gal, D. & Rucker, D. D. (2018). The loss of loss aversion: Will it loom larger than its gain? *Journal of Consumer Psychology*, 28(3), 497–516.

Gill, D. & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1), 469–503.

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, 38(1), 79–109.

Heffetz, O. (2021). Are reference points merely lagged beliefs over probabilities? *Journal of Economic Behavior & Organization*, 181(C), 252–269.

Heffetz, O. & List, J. A. (2014). Is the endowment effect an expectations effect? *Journal of the European Economic Association*, 12(5), 1396–1422.

Henderson, D. J. & Parmeter, C. F. (2015). *Applied nonparametric econometrics*. Cambridge University Press.

Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, 69(2), 499–513.

Horowitz, J. L. & Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32(6), pp. 2412–2443.

Horowitz, J. L. & Mammen, E. (2010). Oracle-efficient nonparametric estimation of an additive model with an unknown link function. *Econometric Theory*, FirstView, 1–27.

Hsiaw, A. (2013). Goal-setting and self-control. *Journal of Economic Theory*, 148(2), 601–626.

Hsiaw, A. (2018). Goal bracketing and self-control. *Games and Economic Behavior*, 111, 100–121.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS of single index models. *Journal of Econometrics*, 50, 71–120.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *The Journal of Political Economy*, 98(6), pp. 1325–1348.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–91.

Kőszegi, B. & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4), 1133–1166.

Kőszegi, B. & Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4), 1047–1073.

Lindskog, A., Martinsson, P., & Medhin, H. (2022). Risk-taking and others. *Journal of Risk and Uncertainty*, (pp. 1–21).

Markle, A., Wu, G., White, R., & Sackett, A. (2018). Goals as reference points in marathon running: A novel test of reference dependence. *Journal of Risk and Uncertainty*, 56, 19–50.

Marzilli Ericson, K. M. & Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, 126(4), 1879–1907.

Pope, D. & Schweitzer, M. (2010). Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes. *American Economic Review*, Forthcoming.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5), 1281–1292.

Rees-Jones, A. (2018). Quantifying Loss-Averse Tax Manipulation. *The Review of Economic Studies*, 85(2), 1251–1278.

Schwerter, F. (2013). *Social Reference Points and Risk Taking*. Bonn Econ Discussion Papers 11/2013, University of Bonn, Bonn Graduate School of Economics (BGSE).

Seibold, A. (2021). Reference points for retirement behavior: Evidence from german pension discontinuities. *American Economic Review*, 111(4), 1126–65.

Strack, P. & Taubinsky, D. (2021). Dynamic preference "reversals" and time inconsistency. *Working Paper*.

Thakral, N. & Tô, L. T. (2021). Daily labor supply and adaptive reference points. *American Economic Review*, 111(8), 2417–43.

A Deriving Test Statistics

Roadmap This appendix provides details and derivations related to our test statistics. Subsection A.1 summarizes all notation. Subsection A.2 details the necessary technical assumptions. Subsection A.3 presents our first-stage estimator, and after establishing a long list of necessary intermediate results it provides a proof of the asymptotic distribution of our claimed estimator. Subsection A.4 presents our second-stage test.

A.1 Summary of Notation

i: individuals Throughout the appendix, we will use i to denote an individual who participates in the experiment. When the derivations involve two or more individuals, they will be denoted by i_1, i_2, i_3, ... and so on.

j: choices Throughout the appendix, we will use (i, j) to denote the jth choice made by individual i. When the derivations involve two or more choices from the same individual i, they will be denoted by (i, j), (i, j'), $(i, \tilde{j})$, ... and so on. When the derivations involve two or more choices from two distinct individuals i and i', they will be denoted by (i, j) and (i', j').

Δ: shifter For each individual i and for each choice j, $\Delta_{(i,j)}$ is the shifter associated with the Δ-shifted gambles presented in that choice problem. Note that $\Delta_{(i,j)}$ is identical and independently distributed across all tuples of (i, j).

r: reference point We denote the reference point of person i by r_i. It is identical and independently distributed across i.

ϵ: utility shock The utility shock (formally specified in Assumption 1) for person i in choice j is denoted $\epsilon_{(i,j)}$.

Prediction Error The error term associated with predicting the choice probability of binary choice with a conditional expectation function is denoted by $v_{(i,j)}$ in the 1st stage,

and $u_{(i,j)}$ in the 2nd stage. The estimated prediction error is denoted by the hat version of corresponding variables.

Number of observations Number of individuals is denoted by n; note that each individual is a cluster in our framework. Number of observation in each cluster (i.e. number of choices made by each subject) is denoted by m. Total number of observations, i.e. $n * m$, is denoted by N. When studying asymptotic properties we consider sampling greater numbers of individuals while holding their number of choices fixed—i.e., we treat m as a constant, and as a result N and n have the same order.

Bandwidth The test involves the usage of kernel smoothing to estimate the unrestricted model, i.e. the probability of choosing the risky gamble conditional on Δ and r. We use h to denote the bandwidth. The test also involves the usage of kernel smoothing to estimate the restricted model, i.e. the null hypothesis that the conditional probability can be represented as a single index model $g(\Delta - r)$. In this case, the bandwidth for estimating the function g is denoted by a. Note that the use of a and h is the same as that in Robinson (1988) and Fan & Li (1996).

Functions: Kernel Smoothing and Probability Density Uni-dimensional kernels are denoted by $k(.)$. For the 1st-stage test, in case of multivariate smoothing over (Δ, r), we use the product of $k(.)$ over the individual dimensions, which is denoted by $K^h(.)$. Also, any function that are over the single dimension of $(\Delta - r)$ is indexed by a subscript or superscript a. For example, the smoothing kernel over the dimension of $(\Delta - r)$ is denoted by K^a, and the density of $\Delta_{(i,j)} - r_i$ is denoted by $f_a(\Delta_{(i,j)} - r_i)$. We will use subscripts like $K^h_{(i,j),(i',j')}$ to denote the pair of observations plugged in the kernel $K^h(.)$. Unless otherwise specified, we use $f(.,.)$ or $f(.,.,.)$ to denote the joint probability distribution of several random variables. The notations here largely follows Fan & Li (1996).

Functions: Conditional Correlation and Variance within Cluster When dealing with calculations involving within-subject heterogeneity, in the 1st stage for example, ρ and σ is defined such that:

$$\sigma^2(\Delta_{(i,j)}, r_i) \equiv E[v^2_{(i,j)}|\Delta_{(i,j)}, r_i]$$

$$\sigma^4(\Delta_{(i,j)}, r_i) \equiv (\sigma^2(\Delta_{(i,j)}, r_i))^2$$

$$\rho^2(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i) \equiv E[v_{(i,j)}v_{(i,j')}|\Delta_{(i,j)}, \Delta_{(i,j')}, r_i]$$

$$\rho^4(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i) \equiv (\rho^2(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i))^2$$

Asymptotic order Throughout this section we use conventional symbols for asymptotic order. Specifically, as number of subjects n approaches infinity, the relationship between x and y, which both depend on n are defined as follows:

(I) $x = O(y)$: There exist constants $0 < M < \infty$ and $0 < L < \infty$ such that for any $n > L$, $x < My$.

(II) $x = o(y)$: for any constant $0 < M < \infty$ there exists a constant $0 < L < \infty$ such that for any $n > L$, $x < My$.

(III) $x = O_p(y)$: x and y are random variables. There exist constants $0 < M < \infty$ and $0 < L < \infty$ such that for any $n > L$ and $\tilde{\epsilon} > 0$, $P(x < My) > 1 - \tilde{\epsilon}$.

(IV) $x = o_p(y)$: x and y are random variables. For any constant $0 < M < \infty$ there exists a constant $0 < L < \infty$ such that for any $n > L$ and $\tilde{\epsilon} > 0$, $P(x < My) > 1 - \tilde{\epsilon}$.

A.2 Technical Assumptions

Our non-parametric approach relies on several relatively mild technical assumptions.

Assumption 5. *The random vector* $(\Delta_{(i,1)}, ..., \Delta_{(i,m)}, r_i, \epsilon_{(i,1)}, ..., \epsilon_{(i,m)})$ *has the following properties:*

(I) It is identically and independently distributed across i.

(II) $(\epsilon_{(i,1)}, ..., \epsilon_{(i,m)}) \perp (\Delta_{(i,1)}, ..., \Delta_{(i,m)}, r_i)$.

(III) $(\epsilon_{(i,1)}, ..., \epsilon_{(i,m)})$ *has a continuous distribution.*

Note that (I) and (II) are satisfied in the data generating process underlying our experiment. (III) is an assumption that most applications of discrete choice model uphold.

Following Robinson (1988) and Fan & Li (1996), we need to define the following two classes of functions before presenting the next assumption.

Definition 1. *$\mathcal{K}_l$, $l \geq 1$, is the class of even functions $k : R \to R$ satisfying*

$$\int_R s^\varphi k(s)ds = max\{1-\varphi, 0\}$$

for any $\varphi = 0, 1, ..., l-1$, and there exist $\delta > 0$ such that

$$k(s) = O((1+|s|^{l+1+\delta})^{-1})$$

Definition 2. *$\mathcal{J}_\gamma^\delta$, $\gamma > 0$, $\delta > 0$, is the class of functions $g : R^d \to R$ satisfying the following properties.*
(I) There exists η such that $\eta - 1 < \gamma < \eta$ and g is $\eta - 1$ times differentiable.
(II) There exists $\epsilon > 0$ such that for any z, $sup_{y\in\{y:|y-z|<\epsilon\}}|g(y) - g(z) - Q_g(y,z)|/|y-z|^\gamma \leq h_g(z)$, where $Q_g = 0$ when $\eta = 1$, Q_g is a $(\eta-1)$th degree homogeneous polynomial in $y - z$ with coefficients the partial derivatives of g at z of orders 1 through $\eta - 1$ and less when $\eta > 1$, and $h_g(z)$ have finite δth moments.

Definition 2 involves a slight abuse of notation: it defines the class of functions $\mathcal{J}_\gamma^\delta$ despite the fact that $\mathcal{J}$ was previously used to denote gambles. The notation here is different from Fan & Li (1996) to avoid the abuse of notation $\mathcal{G}$.

The next technical assumption concerns the properties of the kernels and function under null hypothesis using Definition 1 and 2. It follows Assumption A1 in Fan (1996), with minor revisions that accommodate the issue of clustering in our environment:

Assumption 6. *Any kernel in the test belongs to function class $\mathcal{K}_2$, and $f_a \in \mathcal{J}_\xi^\infty$, for some $1 < \xi \leq 2$, $g \in \mathcal{J}_\xi^4$ where $0 < \gamma < 1$*

Observation 1. *The Gaussian kernel satisfies Assumption 6.*

The final technical assumption concerns the bandwidths in kernel smoothing. It follows Assumption A2 in Fan (1996), with minor revisions that accommodate the issue of clustering in our environment:

Assumption 7. *As* $N \to \infty$, $a \to 0$, $h \to 0$, $Nh^2 \to \infty$, $Na^{2\eta}h \to 0$, $h/a \to 0$, $nh \to \infty$, $na^{1+\eta} \to \infty$, $m^2 h \to 0$, *where* $\eta = min(\xi + 1, \gamma)$, *where* ξ *and* γ *are defined in Assumption 6.*

The lemma below is a direct application of Lemma B.1 in Fan & Li (1996) in our setting, taking into account that r is constant within each cluster:

Lemma 2. *If Assumption 6 and Assumption 7 hold, there exists a function* $D_g(\Delta_{(i,j)}, r_i)$ *which has fourth moment, such that as long as* $(i,j) \neq (i',j')$, $E[[g(\Delta_{(i',j')} - r_i') - g(\Delta_{(i,j)} - r_i)]k^a(\frac{\Delta_{(i,j)} - \Delta_{(i',j')} - r_i + r_i'}{a})|\Delta, r] \leq D_g(\Delta_{(i,j)}, r_i)a^{1+\eta}$, *where* η *is defined in Assumption 7.*

Finally, we list the null hypothesis that is formally tested by our proposed estimator.

$$H_0 : \text{There exists a function } g \in \mathcal{J}_\gamma^4 \text{ such that } E[Y|\Delta, r] = g(\Delta - r).$$

A.3 Deriving 1st-Stage Test Statistic

Our statistic of interest is:

$$E[vf(\Delta - r)E[vf(\Delta - r)|\Delta, r]] \tag{13}$$

where $f(\Delta - r)$ is the p.d.f. of $(\Delta - r)$. Thus we need to approximate v, $f(\Delta - r)$, $f(\Delta, r)$ respectively. The estimator that we adopt is essentially the same as Fan & Li (1996), but modified to accommodate the reference point r_i being held constant within each cluster. To that end, we replace Fan and Li's leave-one-out estimator with a leave-m-out estimator. Specifically, when estimating the functional value (say, probability density function of (Δ, r)) evaluated at $(\Delta_{(i,j)}, r_i)$, we will use every observation other than those which are generated by the same subject.

We estimate the test statistic as follows. First, define

$$\hat{g}_{(i,j)} \equiv \widehat{E}[Y_{(i,j)}|\Delta_{(i,j)} - r_i] = \frac{[(N-m)a]^{-1}\sum_{i'\neq i}\sum_{j,j'} Y_{(i',j')}K^a_{(i,j),(i',j')}}{\widehat{f}_a(\Delta_{(i,j)} - r_i)} \tag{14}$$

where

$$\widehat{f}_a(\Delta_{(i,j)} - r_i) = \frac{1}{(N-m)a}\sum_{i'\neq i}\sum_{j,j'} K^a_{(i,j),(i',j')} \tag{15}$$

and $K^a_{(i,j),(i',j')} \equiv k(\frac{(\Delta_{(i,j)}-r_i)-(\Delta_{(i',j')}-r_{i'})}{a})$. $k(.)$ is a univariate Gaussian kernel.
$E[vf(\Delta - r)E[vf(\Delta - r)|\Delta, r]]$ may then be estimated by

$$\Omega = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_j\sum_{j'}[\bar{v}_{(i,j)}\widehat{f}_a(\Delta_{(i,j)} - r_i)][\bar{v}_{(i',j')}\widehat{f}_a(\Delta_{(i',j')} - r_{i'})]K_{(i,j),(i',j')} \tag{16}$$

where

$$\bar{v}_{(i,j)} = Y_{(i,j)} - \widehat{E}[Y_{(i,j)}|\Delta_{(i,j)} - r_i] \tag{17}$$

$K_{(i,j),(i',j')} \equiv k(\frac{\Delta_{(i,j)}-\Delta_{(i',j')}}{h})k(\frac{r_i-r_j}{h})$ is a product of two univariate Gaussian kernels, and the bandwidth is h.

The statistic Ω is the key element of the test, as elaborated in the theorem below:

Theorem 1. *When assumption 5, 6, and 7 hold:*

(I) Under the null, $Nh\Omega \to N(0, 2(\sigma_a^2 + \rho_a^2))$, *where*

$$\sigma_a^2 = E[f(\Delta_{(i,j)}, r_i)\sigma^4(\Delta_{(i,j)}, r_i)f_a^4(\Delta_{(i,j)} - r_i)][\int k^2(s)ds]^2 \tag{18}$$

and

$$\rho_a^2 = (m^2-1)h(E[(\rho^4(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i))f(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i)f_a^4(\Delta_{(i,j)} - r_i)]\int k^2(s)ds \tag{19}$$

(II) Under the alternative, $Nh\Omega$ *converges to positive infinity with probability 1.*

In light of the theorem, the final test statistic is:

$$T_1 = \frac{Nh\Omega}{\sqrt{2(\hat{\sigma}_a^2 + \hat{\rho}_a^2)}} \tag{20}$$

The estimator $\hat{\sigma}_a^2$, resembling that in Fan & Li (1996), is

$$\hat{\sigma}_a^2 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} [\bar{v}_{(i,j)} \widehat{f}_a(\Delta_{(i,j)} - r_i)]^2 [\bar{v}_{(i',j')} \widehat{f}_a(\Delta_{(i',j')} - r_{i'})]^2 K_{(i,j),(i',j')} [\int k^2(s) ds]^2 \tag{21}$$

As we need to correct for the presence of within-cluster correlation, we need to consistently estimate the second term ρ_a^2. Since we can rewrite ρ_a^2 as

$$\rho_a^2 = (m^2-1)hE[v_{(i,j_1)} f_a(\Delta_{(i,j_1)} - r_i) v_{(i,j_2)} f_a(\Delta_{(i,j_2)} - r_i)$$
$$E[v_{(i',j_1')} f_a(\Delta_{(i',j_1')} - r_{i'}) v_{(i',j_2')} f_a(\Delta_{(i',j_2')} - r_{i'}) | \Delta_{(i',j_1')}, \Delta_{(i',j_2')}, r_{i'}] f(\Delta_{(i',j_1')}, \Delta_{(i',j_2')}, r_{i'})] \int k^2(s) ds \tag{22}$$

This motivates us to use the following U-statistic to estimate it:

$$\hat{\rho}_a^2 = \frac{(m^2-1)h}{N(N-m)(m-1)^2 h^3} \sum_i \sum_{j_1 \neq j_2} \hat{v}_{i,j_1} \hat{f}_a(\Delta_{(i,j_1)} - r_i) \hat{v}_{(i,j_2)} \hat{f}_a(\Delta_{(i,j_2)} - r_i) \sum_{i' \neq i} \sum_{j_1' \neq j_2'}$$
$$\hat{v}_{(i',j_1')} \hat{f}_a(\Delta_{(i',j_1')} - r_{i'}) \hat{v}_{(i',j_2')} \hat{f}_a(\Delta_{(i',j_2')} - r_{i'}) k(\frac{\Delta_{(i,j_1)} - \Delta_{(i,j_2)}}{h}) k(\frac{\Delta_{(i',j_1')} - \Delta_{(i',j_2')}}{h}) k(\frac{r_{i'} - r_i}{h}) \int k^2(s) ds \tag{23}$$

where K is the product kernel of univariate Gaussian kernel $k(.)$.

A.3.1 Intermediate Results for Proof of Theorem 1

To characterize the asymptotic distribution of Ω, we first decompose it into six parts (similar to in Fan & Li (1996), equation (A.1)). We will then present intermediate results characterizing each of these parts before combining results into the final proof of Theorem 1.

To simplify presentation, we will extensively use the following short-hand notation: $f_{a_{(i,j)}} \equiv f_a(\Delta_{(i,j)} - r_i)$, $g_{(i,j)} \equiv g(\Delta_{(i,j)} - r_i)$.

The decomposition is

$$\begin{aligned}\Omega = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}\{&(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i,j)}}(g_{(i',j')} - \hat{g}_{(i',j')})\hat{f}_{a_{(i',j')}} + v_{(i,j)}v_{(i',j')}\hat{f}_{a_{(i,j)}}\hat{f}_{a_{(i',j')}} \\ &+\hat{v}_{(i,j)}\hat{v}_{(i',j')}\hat{f}_{a_{(i,j)}}\hat{f}_{a_{(i',j')}} + 2v_{(i,j)}\hat{f}_{a_{(i,j)}}(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i',j')}} \\ &-2\hat{v}_{(i,j)}\hat{f}_{a_{(i,j)}}(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i',j')}} - 2v_{(i,j)}\hat{f}_{a_{(i,j)}}\hat{v}_{(i',j')}\hat{f}_{a_{(i',j')}}\}K_{(i,j),(i',j')} \\ &\equiv \omega_1 + \omega_2 + \omega_3 + 2\omega_4 - 2\omega_5 - 2\omega_6\end{aligned} \tag{24}$$

where

$$\omega_1 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i,j)}}(g_{(i',j')} - \hat{g}_{(i',j')})\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

$$\omega_2 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}v_{(i,j)}v_{(i',j')}\hat{f}_{a_{(i,j)}}\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

$$\omega_3 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}\hat{v}_{(i,j)}\hat{v}_{(i',j')}\hat{f}_{a_{(i,j)}}\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

$$\omega_4 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}v_{(i,j)}\hat{f}_{a_{(i,j)}}(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

$$\omega_5 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}\hat{v}_{(i,j)}\hat{f}_{a_{(i,j)}}(g_{(i,j)} - \hat{g}_{(i,j)})\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

$$\omega_6 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_{j,j'}v_{(i,j)}\hat{f}_{a_{(i,j)}}\hat{v}_{(i',j')}\hat{f}_{a_{(i',j')}}K_{(i,j),(i',j')}$$

The strategy of our proof will be to establish that, under the null, ω_2 has a known asymptotic distribution and ω_1, ω_3, ..., ω_6 all are asymptotically negligible. The following

propositions establish those claims:

Proposition 7. $\omega_1 = o_p((Nh)^{-1})$

Proof. As discussed in Fan & Li (1996) Proposition A.1, it suffices to show that $E[\omega_1] = o((Nh)^{-1})$ and $E[\omega_1^2] = o((Nh)^{-2})$. From equation 24 we know that

$$\omega_1 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{a_{(i,j)}} (g_{(i',j')} - \hat{g}_{(i',j')}) \hat{f}_{a_{(i',j')}} K_{(i,j),(i',j')}$$

$$= \frac{1}{N(N-m)^3 h^2 a^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{\tilde{j},\tilde{j}'} (g_{(i,j)} - g_{(\tilde{i},\tilde{j})}) K^a_{(i,j),(\tilde{i},\tilde{j})} (g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')}) K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')} \tag{25}$$

The proof of elements where i, i', $\tilde{i}$, $\tilde{i}'$ do not equal each other is exactly the same as Proposition A.1 in Fan & Li (1996). We need to ensure that the within-cluster interaction does not change the original conclusion.

Showing $E[\omega_1] = o((Nh)^{-1})$ The sum of terms where exactly two among i, i', $\tilde{i}$, $\tilde{i}'$ are the same depends on the number of such terms ($n(n-1)^2m^4$), as well as whether the value of kernel function decreases in the order of smoothing bandwidth. In this case, each term does not exceed the order of ah^2. Thus the sum is

$$O(\frac{1}{N(N-m)^3h^2a^2} * n(n-1)^2m^4 * ah^2) = O((na)^{-1}) = o((Nh)^{-1})$$

Showing $E[\omega_1^2] = o((Nh)^{-2})$ We have

$$E[\omega_1^2] = \frac{1}{N^2(N-m)^6h^4a^4} \sum_i \sum_{i' \neq i} \sum_{j,j'} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{\tilde{j},\tilde{j}'} \sum_k \sum_{k' \neq k} \sum_{l,l'} \sum_{\tilde{k} \neq k} \sum_{\tilde{k}' \neq k'} \sum_{\tilde{l},\tilde{l}'}$$
$$(g_{(i,j)} - g_{(\tilde{i},\tilde{j})}) K^a_{(i,j),(\tilde{i},\tilde{j})} (g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')}) K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')}$$
$$(g_{(k,l)} - g_{(\tilde{k},\tilde{l})}) K^a_{(k,l),(\tilde{k},\tilde{l})} (g_{(k',l')} - g_{(\tilde{k}',\tilde{l}')}) K^a_{(k',l'),(\tilde{k}',\tilde{l}')} K_{(k,l),(k',l')}$$

When i, i', $\tilde{i}$, $\tilde{i}'$, k, k', $\tilde{k}$, $\tilde{k}'$ do not equal each other, the expression can be dissected into independent pieces like $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})}) K^a_{(i,j),(\tilde{i},\tilde{j})}$ so that Lemma 2 can be applied. In this case

the sum of these terms is $O(a^{4\eta}) = o((nh)^{-2})$

When exactly two i-indices equal each other, three types of terms need to be discussed.

(i) $i = k$. In this case, conditional on i, i', k, k', Lemma 2 can be applied to all terms like $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^a_{(i,j),(\tilde{i},\tilde{j})}$, so that the product of these conditional expectation terms is of order $O(a^{4+4\eta})$. As $i = k$, $K_{(i,j),(i',j')} * K_{(k,l),(k',l')}$ is of order $O(h^3)$. Therefore the sum of these terms is $\frac{1}{nh} * O(a^{4\eta}) = o((nh)^{-2})$

(ii) $i = \tilde{k}$. In this case, conditional on i, i', k, k', Lemma 2 can be applied to all terms like $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^a_{(i,j),(\tilde{i},\tilde{j})}$ except for $(g_{(k,l)} - g_{(\tilde{k},\tilde{l})})K^a_{(k,l),(\tilde{k},\tilde{l})}$, so that the product of these conditional expectation terms is of order $O(a^{3+3\eta})$. Therefore the sum of these terms is $\frac{1}{n} * O(a^{3\eta}) = o(n^{-2}h^{-1}a^{\eta})) = o((nh)^{-2})$

(iii) $\tilde{i} = \tilde{k}$, this case is similar to (i). In this case, conditional on i, i', k, k', $\tilde{i}$, Lemma 2 can be applied to $(g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')})K^a_{(i',j'),(\tilde{i}',\tilde{j}')}$ and $(g_{(k,l)} - g_{(\tilde{k},\tilde{l})})K^a_{(k,l),(\tilde{k},\tilde{l})}(g_{(k',l')} - g_{(\tilde{k}',\tilde{l}')})K^a_{(k',l'),(\tilde{k}',\tilde{l}')}$, whose conditional expectation is both of order $O(a^{1+\eta})$. Then, conditional on i, i', k, k', the expectation of order $O(a^{2+2\eta})$. The order in total is thus $\frac{1}{n}O(a^{4\eta}) = o((nh)^{-2})$

When i, i', $\tilde{i}$, $\tilde{i}'$, k, k', $\tilde{k}$, $\tilde{k}'$ takes no more than six values, the order is at most $max\{O(n^{-2} * \frac{1}{h^2} * a^{4\eta}), O((na)^{-2}a^{2\eta})\} = o((nh)^{-2})$

□

Proposition 8. *$Nh\omega_2 \to N(0, 2(\sigma_a^2 + \rho_a^2))$ in distribution, where*

$$\sigma_a^2 = E[f(\Delta_{(i,j)}, r_i)\sigma^4(\Delta_{(i,j)}, r_i)f_a^4(\Delta_{(i,j)} - r_i)][\int k^2(s)ds]^2$$

and

$$\rho_a^2 = (m^2 - 1)h(E[(\rho^4(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i))f(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i)f_a^4(\Delta_{(i,j)} - r_i)] \int k^2(s)ds$$

where k(.) is the Gaussian kernel.

Proof. Consider term ω_2. We have

$$\begin{aligned}\omega_2 &= \frac{1}{N(N-m)^3h^2a^2}\sum_i\sum_{i'\neq i}\sum_{\tilde{i}\neq i}\sum_{\tilde{i}'\neq i'}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')} \\ &\equiv \frac{1}{N(N-m)^3h^2a^2}\sum_{i\neq i'\neq\tilde{i}\neq\tilde{i}'}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')} \\ &\quad +\omega_2^R \\ &\equiv \omega_2^U + \omega_2^R\end{aligned} \tag{26}$$

Here the terms where i, i', $\tilde{i}$, $\tilde{i}'$ do not equal each other is denoted by ω_2^U. This is a key place in the proof where the leave-one-out estimator in Fan & Li (1996) needs to be revised, since we cannot rely on the independence of different observations within the same cluster to eliminate some relevant cross-products. Our leave-m-out estimator, where the cross products of some dependent observations are omitted, addresses this issue.

To see this formally, rewrite ω_2^U in terms of U-statistics:

$$\frac{\binom{n}{4}}{N(N-m)^3h^2a^2}[\binom{n}{4}^{-1}\sum_{1\leq i<i'<\tilde{i}<\tilde{i}'\leq n} P(\mathcal{Z}_i,\mathcal{Z}_{i'},\mathcal{Z}_{\tilde{i}},\mathcal{Z}_{\tilde{i}'})]$$

where

$$\mathcal{Z}_i = (\Delta_{(i,1)},...,\Delta_{(i,m)},v_{(i,1)},...,v_{(i,m)},r_i)'$$

and

$$P(\mathcal{Z}_i,\mathcal{Z}_{i'},\mathcal{Z}_{\tilde{i}},\mathcal{Z}_{\tilde{i}'}) = \sum_{4!}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')}$$

where 4! stands for the permutation of $\{i,i',\tilde{i},\tilde{i}'\}$

Define $P_n(\mathcal{Z}_i,\mathcal{Z}_{i'}) = E[P(\mathcal{Z}_i,\mathcal{Z}_{i'},\mathcal{Z}_{\tilde{i}},\mathcal{Z}_{\tilde{i}'})|\mathcal{Z}_i,\mathcal{Z}_{i'}]$, we have

$P_n(\mathcal{Z}_i,\mathcal{Z}_{i'}) = 4\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K_{(i,j),(i',j')}E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}]$

$$E[P_n(\mathcal{Z}_i,\mathcal{Z}_{i'})^2] = 16E[(\sum_{j,j'} v_{(i,j)}v_{(i',j')}K_{(i,j),(i',j')}\sum_{\tilde{j},\tilde{j}'} E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]$$

$$= 16m^4E[(\sum_{j,j'} v_{(i,j)}v_{(i',j')}K_{(i,j),(i',j')}E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]$$

$$= 16m^6E[v^2_{(i,j)}v^2_{(i',j')}K^2_{(i,j),(i',j')}(E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]+$$

$$16m^6(m^2-1)E[v_{(i,j)}v_{(i',j')}v_{(i,j_*)}v_{(i',j'_*)}K_{(i,j),(i',j')}K_{(i,j_*),(i',j'_*)}(E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2] \tag{27}$$

The first term, as the original proof derives, can be reduced to

$16m^6a^4h^2E[f(\Delta_{(i,j)},r_i)\sigma^4(\Delta_{(i,j)},r_i)f_a^4][\int k^2(s)ds]^2$

Next simplify the second term. When a and h is approaching 0, we have:

$$16m^6(m^2-1)E[v_{(i,j)}v_{(i',j')}v_{(i,j_*)}v_{(i',j'_*)}K_{(i,j),(i',j')}K_{(i,j_*),(i',j'_*)}(E[K^a_{(i,j),(\tilde{i},\tilde{j})}K^a_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}]))^2]$$

$$= 16m^6(m^2-1)a^4E[\rho^2(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)\rho^2(\Delta_{(i',j')},\Delta_{(i',j'_*)},r_{i'})K_{(i,j),(i,j_*)}K_{(i',j'),(i',j'_*)}$$

$$(\int k(\tau_{(i,j),(\tilde{i},\tilde{j})})k(\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})f_a(\Delta_{(i,j)}-r_i+a\tau_{(i,j),(\tilde{i},\tilde{j})})$$

$$f_a(\Delta_{(i',j')}-r_{i'}+a\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})d\tau_{(i,j),(\tilde{i},\tilde{j})}d\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})^2]$$

$$= 16m^6(m^2-1)a^4h^3\int \rho^2(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)\rho^2(\Delta_{(i,j)}+hs_1,\Delta_{(i,j_*)}+hs_2,r_i+hs_3)f(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)$$

$$f(\Delta_{(i,j)}+hs_1,\Delta_{(i,j_*)}+hs_2,r_i+hs_3)k(s_1)k(s_2)k^2(s_3)$$

$$(\int k^a(\tau_{(i,j),(\tilde{i},\tilde{j})})k^a(\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})f_a(\Delta_{(i,j)}-r_i+a\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})$$

$$f_a(\Delta_{(i,j)}+hs_1-r_i-hs_2+a\tau_{(i,j),(\tilde{i},\tilde{j})})d\tau_{(i,j),(\tilde{i},\tilde{j})}d\zeta_{(i',j'),(\tilde{i}',\tilde{j}')})^2ds_1ds_2ds_3d\Delta_{(i,j)}d\Delta_{(i,j_*)}dr_i$$

$$= 16m^6(m^2-1)a^4h^3(E[(\rho^4(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i))f(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)f_a(\Delta_{(i,j)}-r_i)^4]\int k^2(s_3)ds_3+o(1))$$

where $s_1 \equiv \frac{\Delta_{i',j'}-\Delta_{i,j}}{h}$, $s_2 \equiv \frac{\Delta_{i,j'_*}-\Delta_{i,j_*}}{h}$, $s_3 \equiv \frac{r_{i'}-r_i}{h}$, $\tau_{(i,j),(\tilde{i},\tilde{j})} \equiv \frac{(\Delta_{(\tilde{i},\tilde{j})}-r_{\tilde{i}})-(\Delta_{(i,j)}-r_i)}{a}$, $\zeta_{(i',j'),(\tilde{i}',\tilde{j}')} \equiv \frac{(\Delta_{(\tilde{i}',\tilde{j}')}-r_{\tilde{i}'})-(\Delta_{(i',j')}-r_{i'})}{a}$, and $f(.,.,.)$ is the joint probability distribution of $(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)$.

Thus when within-cluster correlation is taken into account, we have

$$Nh\omega_2^U \xrightarrow{d} N(0,2(\sigma_a^2+\rho_a^2))$$

where

$$\rho_a^2 = (m^2-1)h(E[(\rho^4(\Delta_{(i,j)},\Delta_{(i,j')},r_i))f(\Delta_{(i,j)},\Delta_{(i,j')},r_i)f_a^4(\Delta_{(i,j)}-r_i)]\int k^2(s)ds$$

Finally we have,

$$\begin{aligned}\omega_2^R = \frac{1}{N(N-m)^3h^2a^2}\{&\sum_{i\neq i'\neq\tilde{i}}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i},\tilde{j}')}K_{(i,j),(i',j')}+\\ &\sum_{i\neq i'\neq\tilde{i}}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^w_{(i,j),(i',\tilde{j})}K^w_{(i',j'),(\tilde{i},\tilde{j}')}K_{(i,j),(i',j')}+\\ &\sum_{i\neq j}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^w_{(i,j),(i',\tilde{j})}K^w_{(i',j'),(i,\tilde{j}')}K_{(i,j),(i',j')}\}\\ &\equiv \omega_2^{R_1}+\omega_2^{R_2}+\omega_2^{R_3}\end{aligned}$$

There are three terms in total. For the first term $\omega_2^{R_1}$ and second term $\omega_2^{R_2}$, we can use similar arguments deriving the distribution of degenerate U-statistics to estimate the order of its second moment. For the third term $\omega_2^{R_3}$, directly estimate its order using LLN to derive the order of its second moment. Then we would have,

$$\begin{aligned}Nh\omega_2^{R_1} &= \frac{\binom{n}{3}}{(N-m)^3ha^2}[\binom{n}{3}^{-1}\omega_2^{R_1}]\\ &= n^{-1}\frac{\binom{n}{3}}{(N-m)^3ha^2}[n\binom{n}{3}^{-1}\omega_2^{R_1}]\end{aligned}$$

As m is a constant, and $n\binom{n}{3}^{-1}\omega_2^{R_1} = o_p(ah^2)$, we have

$$Nh\omega_2^{R_1} = o_p(\frac{1}{na}) = o_p(1)$$

Similarly we have

$$Nh\omega_2^{R_2} = o_p(\frac{1}{na}) = o_p(1)$$

and

$$Nh\omega_2^{R_3} = o_p(\frac{1}{na}) = o_p(1)$$

So now we can conclude that

$$Nh\omega_2 \to N(0, 2(\sigma_a^2 + \rho_a^2))$$

□

Proposition 9. $\omega_3 = o_p((Nh)^{-1})$

Proof. From equation 24, we have,

$$\omega_3 = \frac{1}{N(N-m)^3h^2a^2} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{j,j',\tilde{j},\tilde{j}'} v_{(\tilde{i},\tilde{j})} v_{(\tilde{i}',\tilde{j}')} K^a_{(i,j),(\tilde{i},\tilde{j})} K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')}$$
$$\equiv \omega_3^F + \omega_3^S$$

where ω_3^F is the sum of the terms where $\tilde{i} = \tilde{i}'$. ω_3^S is the sum of the terms where $\tilde{i} \neq \tilde{i}'$.

For ω_3^F, we have

$$E[\omega_3^F] = \frac{1}{N(N-m)^3h^2a^2} * O(\sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{j,j',\tilde{j},\tilde{j}'} E[v_{(\tilde{i},\tilde{j})} v_{(\tilde{i},\tilde{j}')} K^a_{(i,j),(\tilde{i},\tilde{j})} K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')}])$$
$$= O(\frac{1}{na}) = o(\frac{1}{nh})$$

and

$$E[(\omega_3^F)^2] = o(\frac{1}{(na)^2}) = o(\frac{1}{(nh)^2})$$

$$\omega_3^S = \omega_3^{SF} + \omega_3^{SS}$$

where ω_3^{SF} is the sum of the terms where i, i', $\tilde{i}$, $\tilde{i}'$ take different values. ω_3^{SS} is the sum of the terms where at least two of the values equal each other.

Regarding ω_3^{SF}, we have,

$$\begin{aligned} E[(\omega_3^{SF})^2] &= \frac{1}{n^8h^4a^4} * o(n^6h^4 * a * 2) * max\{o(\frac{1}{nh}, \frac{1}{na}, 1)\} \\ &= o(\frac{1}{n^2a^2}) = o(\frac{1}{n^2h^2}) \end{aligned}$$

Regarding ω_3^{SS}, we have,

$$E[|\omega_3^{SS}|] = \frac{1}{N^4h^2a^2} * O(N^3ah^2) = O(\frac{1}{na}) = o(\frac{1}{nh})$$

□

Proposition 10. $\omega_4 = o_p((Nh)^{-1})$

Proof. We have

$$\omega_4 = \sum_i \sum_{i' \neq i} v_{(i,j)} \sum j, j' \hat{f}_{a_{(i,j)}}(g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{a_{(i',j')}}$$

$$= \frac{1}{N(N-m)^3h^2a^2} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum j, j', \tilde{j}, \tilde{j}' v_{(i,j)}(g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')}) K^a_{(i,j),(\tilde{i},\tilde{j})} K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')}$$

Since

$$E[\omega_4] = 0$$

and

$$E[\omega_4^2] = \frac{1}{N^2(N-m)^6h^4a^4} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_k \sum_{k' \neq k} \sum_{\tilde{k} \neq k} \sum_{\tilde{k}' \neq k'} \sum_{j,j',\tilde{j},\tilde{j}',l,l',\tilde{l},\tilde{l}'}$$

$$v_{(i,j)}(g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')}) K^a_{(i,j),(\tilde{i},\tilde{j})} K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')} v_{(k,l)}(g_{(k',l')} - g_{(\tilde{k}',\tilde{l}')}) K^a_{(k,l),(\tilde{k},\tilde{l})} K^a_{(k',l'),(\tilde{k}',\tilde{l}')} K_{(k,l),(k',l')}$$

$$\equiv \omega_4^P + \omega_4^R$$

where ω_4^P is the sum of the terms where $i = k$ and $i, i', \tilde{i}, \tilde{i}', k', \tilde{k}, \tilde{k}'$ are pairwise different. ω_4^R represents all the other terms. ω_4^P is of order $O(n^{-1}a^{2\eta}) = o(\frac{1}{n^2h}) = o(\frac{1}{n^2h^2})$. When two of these indices equal each other, ω_4^R is at most of order $O(n^{-1}a^{2\eta}max\{\frac{a}{na^\eta}, \frac{1}{na^\eta}, \frac{1}{nh^2}\}) = o(\frac{1}{n^2h^2})$

□

Proposition 11. $\omega_5 = o_p((Nh)^{-1})$

Proof. As ω_5 is algebraically of the same structure as ω_4, the derivation is the same as Proposition 10. □

Proposition 12. $\omega_6 = o_p((Nh)^{-1})$

Proof. We have:

$$\omega_6 = \sum_i \sum_{i' \neq i} \sum_{j,j'} v_{(i,j)} \hat{f}_{a_{(i,j)}} \hat{v}_{(i',j')} \hat{f}_{a_{(i',j')}} K_{(i,j),(i',j')}$$

$$= \frac{1}{N(N-m)^3h^2a^2} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)} v_{(\tilde{i}',\tilde{j}')} K^a_{(i,j),(\tilde{i},\tilde{j})} K^a_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')}$$

$$\equiv \omega_6^F + \omega_6^S$$

where ω_6^F denotes the sum the of terms where $\tilde{i}' = i$, ω_6^S denotes the sum of the terms where $\tilde{i}' \neq i$. We have, Similar to Proposition A.6 in Fan & Li (1996), we have $E[(\omega_6^F)^2] = o(n^{-2}h^{-2})$ and $E[(\omega_6^S)^2] = o(n^{-2}h^{-2})$. Thus $\omega_6 = o_p((Nh)^{-1})$. □

Proposition 13. $\hat{\sigma}_a^2 = \sigma_a^2 + o_p(1)$ *and* $\hat{\rho}_a^2 = \rho_a^2 + o_p(1)$

Proof. We know from equation 21 that

$$\hat{\sigma}_a^2 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} [\bar{v}_{(i,j)} \widehat{f}_a(\Delta_{(i,j)} - r_i)]^2 [\bar{v}_{(i',j')} \widehat{f}_a(\Delta_{(i',j')} - r_{i'})]^2 K_{(i,j),(i',j')} \int K^2(s) ds$$

using Lemma 2 and discussions in Proposition 7 to Proposition 12, we have,

$$\hat{\sigma}_a^2 = \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_j\sum_{j'}[\bar{v}_{(i,j)}\widehat{f}_a(\Delta_{(i,j)}-r_i)]^2[\bar{v}_{(i',j')}\widehat{f}_a(\Delta_{(i',j')}-r_{i'})]^2K_{(i,j),(i',j')}\int K^2(s)ds + o_p(1)$$
$$= \frac{1}{N(N-m)h^2}\sum_i\sum_{i'\neq i}\sum_j\sum_{j'}[v_{(i,j)}f_a(\Delta_{(i,j)}-r_i)]^2[v_{(i',j')}f_a(\Delta_{(i',j')}-r_{i'})]^2K_{(i,j),(i',j')}\int K^2(s)ds + o_p(1)$$
$$= \sigma_a^2 + o_p(1)$$

Similarly,

$$\hat{\rho}_a^2 = \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}\hat{v}_{i,j_1}\hat{f}_a(\Delta_{(i,j_1)}-r_i)\hat{v}_{(i,j_2)}\hat{f}_a(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$\hat{v}_{(i',j_1')}\hat{f}_a(\Delta_{(i',j_1')}-r_{i'})\hat{v}_{(i',j_2')}\hat{f}_a(\Delta_{(i',j_2')}-r_{i'})k(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h})k(\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h})k(\frac{r_{i'}-r_i}{h})\int k^2(s)ds$$
$$= \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}v_{i,j_1}\hat{f}_a(\Delta_{(i,j_1)}-r_i)v_{(i,j_2)}\hat{f}_a(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$v_{(i',j_1')}\hat{f}_a(\Delta_{(i',j_1')}-r_{i'})v_{(i',j_2')}\hat{f}_a(\Delta_{(i',j_2')}-r_{i'})k(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h})k(\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h})$$
$$k(\frac{r_{i'}-r_i}{h})\int k^2(s)ds + o_p(1)$$
$$= \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}v_{i,j_1}f_a(\Delta_{(i,j_1)}-r_i)v_{(i,j_2)}f_a(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$v_{(i',j_1')}f_a(\Delta_{(i',j_1')}-r_{i'})v_{(i',j_2')}f_a(\Delta_{(i',j_2')}-r_{i'})k(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h})$$
$$k(\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h})k(\frac{r_{i'}-r_i}{h})\int k^2(s)ds + o_p(1)$$
$$= \hat{\rho}_a^2 = \rho_a^2 + o_p(1) \tag{28}$$

□

A.3.2 Proof of Theorem 1

Armed with the results of Appendix A.3.1, we are now equipped to prove Theorem 1. We begin by restating Theorem 1 for readers' convenience.

Theorem 1. *When assumption 5, 6, and 7 hold:*

(I) Under the null, $Nh\Omega \to N(0, 2(\sigma_a^2 + \rho_a^2))$*, where*

$$\sigma_a^2 = E[f(\Delta_{(i,j)}, r_i)\sigma^4(\Delta_{(i,j)}, r_i)f_a^4(\Delta_{(i,j)} - r_i)][\int k^2(s)ds]^2$$

and

$$\rho_a^2 = (m^2 - 1)h(E[(\rho^4(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i))f(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i)f_a^4(\Delta_{(i,j)} - r_i)] \int k^2(s)ds$$

(II) Under the alternative, $Nh\Omega$ *converges to positive infinity with probability 1.*

Proof. To establish part (I), recall from equation 16 that we can decompose the test statistics into

$$\omega_1 + \omega_2 + \omega_3 + 2\omega_4 - 2\omega_5 - 2\omega_6$$

Proposition 8 establishes that ω_2 has exactly the distribution noted in the theorem. Propositions 7, 9, 10, 11, and 12 establish that all other terms are asymptotically negligible. Proposition 13 establishes the consistency of the necessary input estimators $\hat{\sigma}_a^2$ and $\hat{\rho}_a^2$. These results together imply that the test statistic has the stated distribution, completing the proof of part I. Part II of the proof trivially follows from the logic in Fan & Li (1996) Theorem 3.2. □

A.4 Deriving 2nd-Stage Test Statistic

As we describe in Section 2, our proposed procedure involves a 2nd-stage test to be undertaken after a failure to reject the null in the first stage. The purpose of this 2nd-stage test is to rule out a degenerate form of reference dependence in which the single-index function maps to a constant. We construct this test with a simple analog of our 1st-stage approach.

Formally, we test the null hypothesis

$$H_0 : \text{There exists a real number } \mu, \text{ such that } E[Y|x] = \mu \text{ for any } x.$$

against the alternative that there does not exist such a μ.

We will use the sample average to approximate μ, that is,

$$\hat{\mu} = \frac{1}{N}\sum_i \sum_j Y_{(i,j)} \tag{29}$$

The estimated approximation error is

$$\hat{u}_{(i,j)} \equiv Y_{(i,j)} - \hat{\mu} \tag{30}$$

The numerator of the test statistic, Π, under the null, is

$$\begin{aligned}
\Pi &= \frac{1}{N(N-m)a}\sum_i \sum_{i'\neq i} \sum_{j,j'} (Y_{(i,j)} - \hat{\mu})(Y_{(i',j')} - \hat{\mu}) k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \\
&= \frac{1}{N(N-m)a}\sum_i \sum_{i'\neq i} \sum_{j,j'} (u_{(i,j)} + \mu - \hat{\mu})(u_{(i',j')} + \mu - \hat{\mu}) k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \\
&= (\mu - \hat{\mu})^2 \frac{1}{N(N-m)a}\sum_i \sum_{i'\neq i} \sum_{j,j'} k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \\
&+ (\mu - \hat{\mu})\frac{1}{N}\sum_i \sum_j u_{(i,j)} \frac{1}{(N-m)a}\sum_{i'\neq i} \sum_{j'} k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \\
&+ (\mu - \hat{\mu})\frac{1}{N}\sum_{i'} \sum_{j'} u_{(i',j')} \frac{1}{(N-m)a}\sum_{i\neq i'} \sum_{j} k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \\
&+ \frac{1}{N(N-m)a}\sum_i \sum_{i'\neq i} \sum_{j,j'} u_{(i,j)} u_{(i',j')} k(\frac{X_{(i',j')} - X_{(i,j)}}{a})
\end{aligned} \tag{31}$$

Since $(\mu - \hat{\mu}) = o(\frac{1}{\sqrt{N}})$, the first three terms are all are all $O_p(\frac{1}{N}) = o_p((N\sqrt{a})^{-1})$ while the last term is $O_p((N\sqrt{a})^{-1})$, we could rescale the term so that the theorem for degenerate U-statistics can again be applied.

$$\begin{aligned}
N\sqrt{a}\Pi &= \frac{N\sqrt{a}}{N(N-m)a}\sum_i \sum_{i'\neq i} \sum_j \sum_{j'} u_{(i,j)} u_{(i',j')} k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) + o_p(1) \\
&= \frac{\binom{n}{2} m}{N(N-m)\sqrt{a}} \{ n \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} [2 \sum_j \sum_{j'} u_{(i,j)} u_{(i',j')} k(\frac{X_{(i',j')} - X_{(i,j)}}{a})] \} + o_p(1) \\
&\to N(0, 2(\sigma_\mu^2 + \rho_\mu^2))
\end{aligned} \tag{32}$$

where

$$\sigma_\mu^2 = E[u^2]^2 E[f(X)] \int k^2(u)du \tag{33}$$

$$\rho_\mu^2 = E[\rho^4(X_{(i,j)}, X_{(i',j')}) f(X_{(i,j)}, X_{(i',j')})] \tag{34}$$

Thus, similar to the first-stage test, we have:

Theorem 2. *(I) Under the null,* $N\sqrt{a}\Pi \to N(0, 2(\sigma_\mu^2 + \rho_\mu^2))$*, where*

$$\sigma_\mu^2 = E[u^2]^2 E[f(X)] \int k^2(u)du$$

$$\rho_\mu^2 = E[\rho^4(X_{(i,j)}, X_{(i,j_*)}) f(X_{(i,j)}, X_{(i,j_*)})]$$

(II) Under the alternative, $N\sqrt{a}\Pi$ *converges to positive infinity with probability 1.*

The estimator for σ_a^2 and ρ_a^2 is

$$\hat{\sigma}_\mu^2 = \frac{1}{N(N-m)a} \sum_i \sum_j \hat{u}_{(i,j)}^2 \sum_{i' \neq i} \sum_{j'} \hat{u}_{(i',j')}^2 k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \int k^2(u)du \tag{35}$$

$$\begin{aligned} \hat{\rho}_\mu^2 = \frac{(m^2-1)a}{N(N-m)(m-1)^2 a^2} \sum_i \sum_{j_1 \neq j_2} \hat{u}_{(i,j_1)} \hat{u}_{(i,j_2)} \sum_{i' \neq i} \sum_{j_1' \neq j_2'} \hat{u}_{(i',j_1')} \hat{u}_{(i',j_2')} \\ k(\frac{X_{(i',j_1')} - X_{(i,j_1)}}{a}) k(\frac{X_{(i',j_2')} - X_{(i,j_2)}}{a}) \end{aligned} \tag{36}$$

The test statistics is

$$T_2 = \frac{N\sqrt{a}\Pi}{\sqrt{2(\hat{\sigma}_\mu^2 + \hat{\rho}_\mu^2)}} \tag{37}$$

Theorem 2 can be proved using the same approach employed in Theorem 1. While still somewhat laborious, it is substantially simpler because of the reduced complexity of the null hypothesis.

B Extending Approach to the Kőszegi-Rabin Framework

The approach presented in our main text addresses a case where a reference point is, in fact, a point, and this point can be exogenously manipulated. These assumptions are satisfied by most common models of reference points, but are notably not satisfied in the model of Kőszegi & Rabin (2006) (or its extension to risk attitudes in Kőszegi & Rabin (2007)). In those frameworks, reference effects are assumed to operate by individuals considering a distribution of different possible reference points, with the distribution pinned down by the distribution of possible consumption realizations given one's choices. The differences of this model make it challenging to nest it within an approach designed for other common reference points. Despite its differences, however, this model can be tested with relatively minor modifications to our existing framework.

To begin, consider the framework for utility as presented in Kőszegi & Rabin (2007). For a riskless wealth outcome $w \in \mathbb{R}$ and riskless reference level of wealth $r \in \mathbb{R}$, define utility as $u(w|r) = m(w) + \mu((m(w) - m(r))$. The term $m(w)$ is viewed as intrinsic "consumption utility," and is analogous to the $\psi(c)$ term in our framework that we refer to as direct utility. The term $\mu(\cdot)$ is the reference-dependent evaluation, analogous to the $\phi(\cdot)$ term in our framework. Note that the input to the reference-dependent evaluation is the difference in consumption utilities derived from wealth outcome w and the reference level, as opposed to the difference in w and r themselves (as in our theory). However, also note that this distinction is inconsequential when consumption utility is assumed to be linear, as we previously imposed in Assumption 2 and as we will continue to impose here.

When extending this concept to potentially risky outcomes, utility is defined by $U(F|G) = \int \int u(w|r)dG(r)dF(w)$, where G is a probability measure over reference points and F is a probability measure over potential wealth outcomes.

To focus ideas, consider Kőszegi and Rabin's solution concept of "choice-acclimating personal equilibrium."

Definition 3. *For any choice set D, $F \in D$ is a choice-acclimating personal equilibrium (CPE) if $U(F|F) \geq U(F'|F')$ for all $F' \in D$. (Kőszegi & Rabin, 2007, Definition 3)*

We now consider how to apply this concept in a framework like our own. In the approach

of Section 1.2, the choice sets presented to subjects always contain two gambles, denoted $\mathcal{G}_0$ and $\mathcal{G}_1$. These gambles were defined over discrete sets of outcomes, $o \in O$ (i.e., heads or tails coin flips as in our experiment). With this set up, we formally define a notion of choice guided by CPE that incorporates additional assumptions that we have imposed.

Assumption 8. *For a choice between two gambles $\mathcal{G}_0$ and $\mathcal{G}_1$, $\mathcal{G}_1$ is chosen if*

$$\sum_{o \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} c_o^{\mathcal{G}_1} + \sum_{o,o' \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} p_{o'}^{\mathcal{G}_1} \mu(c_o^{\mathcal{G}_1} - c_{o'}^{\mathcal{G}_1}) + \epsilon_1 \geq \sum_{o \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} c_o^{\mathcal{G}_0} + \sum_{o,o' \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} p_{o'}^{\mathcal{G}_0} \mu(c_o^{\mathcal{G}_0} - c_{o'}^{\mathcal{G}_0}) + \epsilon_0.$$

Decisions made according to Assumption 8 can be understood as CPE while imposing the assumption that consumption utility m is linear and while introducing a random-utility component ϵ.

We will now show that a single-index structure analogous to that in our primary test can arise under Assumption 8 for a particular class of gambles. Define the *double Δ-shift* operation over base gamble $\mathcal{G}$ and partition of the set of outcomes $O^{\mathcal{G}} = O_a^{\mathcal{G}} \cup O_b^{\mathcal{G}}$ (held constant across both base gambles) to be:

$$S(\Delta_a, \Delta_b | \mathcal{G}, O_a^{\mathcal{G}}, O_b^{\mathcal{G}}) = (p_o^{\mathcal{G}}, c_o^{\mathcal{G}} + \Delta_a I(o \in O_a^{\mathcal{G}}) + \Delta_b I(o \in O_b^{\mathcal{G}}))_{o \in O^{\mathcal{G}}}$$

In this equation, $I(o \in O)$ is an indicator function that takes the value 1 if $o \in O$ and 0 otherwise.

To illustrate with an example, consider the base gambles in base scenario 5 of our experiment, where $G_0 = (100\%, 3.4)$ and $G_1 = (50\%, 2.5; 50\%, 4.5)$. We may form double Δ-shifted gambles by adding Δ_a to each outcome of these two choices and mixing both choices with a payment of Δ_b with a 50% of chance. The revised, double Δ-shifted gamble is $G_0' = (50\%, 3.4 + \Delta_a; 50\%\Delta_b)$ and $G_1' = (25\%, 2.5 + \Delta_a; 25\%, 4.5 + \Delta_a; 50\%, \Delta_b)$.

Similar to the set-up of Proposition 1, consider choice made between double Δ-shifted values of two fixed base gambles. Let Y take the value of 1 if $S(\Delta_a, \Delta_b | \mathcal{G}_1, O_a^{\mathcal{G}_1}, O_b^{\mathcal{G}_1})$ is chosen and the value of 0 if $S(\Delta_a, \Delta_b | \mathcal{G}_0, O_a^{\mathcal{G}_0}, O_b^{\mathcal{G}_0})$ is chosen. If choices are made according to Assumption 8, there exists a single index representation of $E[Y | \Delta_a, \Delta_b]$ as $g(\Delta_a - \Delta_b)$ for

some g. This follows from series of calculations similar to those developed in equations 4-6:

$$
\begin{aligned}
&E[Y|\Delta_a,\Delta_b] = \\
&Pr\Bigg[\sum_{o\in O_a^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}(c_o^{\mathcal{G}_1}+\Delta_a) + \sum_{o\in O_b^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}(c_o^{\mathcal{G}_1}+\Delta_b) && (38)\\
&\quad - \sum_{o\in O_a^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}(c_o^{\mathcal{G}_0}+\Delta_a) - \sum_{o\in O_b^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}(c_o^{\mathcal{G}_0}+\Delta_b) && (39)\\
&+ \sum_{o,o'\in O_a^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}p_{o'}^{\mathcal{G}_1}\mu(c_o^{\mathcal{G}_1}-c_{o'}^{\mathcal{G}_1}) + \sum_{o,o'\in O_b^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}p_{o'}^{\mathcal{G}_1}\mu(c_o^{\mathcal{G}_1}-c_{o'}^{\mathcal{G}_1}) && (40)\\
&- \sum_{o,o'\in O_a^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}p_{o'}^{\mathcal{G}_0}\mu(c_o^{\mathcal{G}_0}-c_{o'}^{\mathcal{G}_0}) - \sum_{o,o'\in O_b^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}p_{o'}^{\mathcal{G}_0}\mu(c_o^{\mathcal{G}_0}-c_{o'}^{\mathcal{G}_0}) && (41)\\
&\quad + \sum_{o\in O_a^{\mathcal{G}_1},o'\in O_b^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}p_{o'}^{\mathcal{G}_1}\mu(c_o^{\mathcal{G}_1}-c_{o'}^{\mathcal{G}_1}+\Delta_a-\Delta_b) && (42)\\
&\quad - \sum_{o\in O_a^{\mathcal{G}_0},o'\in O_b^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}p_{o'}^{\mathcal{G}_0}\mu(c_o^{\mathcal{G}_0}-c_{o'}^{\mathcal{G}_0}+\Delta_a-\Delta_b) && (43)\\
&\quad + \sum_{o\in O_b^{\mathcal{G}_1},o'\in O_a^{\mathcal{G}_1}} p_o^{\mathcal{G}_1}p_{o'}^{\mathcal{G}_1}\mu(c_o^{\mathcal{G}_1}-c_{o'}^{\mathcal{G}_1}+\Delta_b-\Delta_a) && (44)\\
&- \sum_{o\in O_b^{\mathcal{G}_0},o'\in O_a^{\mathcal{G}_0}} p_o^{\mathcal{G}_0}p_{o'}^{\mathcal{G}_0}\mu(c_o^{\mathcal{G}_0}-c_{o'}^{\mathcal{G}_0}+\Delta_b-\Delta_a) \geq \epsilon_0-\epsilon_1\Bigg] && (45)
\end{aligned}
$$

Notice that the terms in lines 40 and 41 solely consist of constants (i.e., they do not include Δ_a and Δ_b), and thus are jointly constant. Denote this constant k_2. Also note that, in all terms on lines 42 through 45, all terms[21] may be expressed as a function of Δ_a-Δ_b, and thus these lines may be collapsed into a single function $\nu(\Delta_a-\Delta_b)$. Finally, note that the differences in consumption utility expressed in lines 38 and 39 will be constant (denoted k_1 with regard to Δ_a and Δ_b if $\sum_{o\in O_a^{\mathcal{G}_0}} p_o = \sum_{o\in O_a^{\mathcal{G}_1}} p_o$. While this is not guaranteed for arbitrary gambles, gambles can be constructed where this is satisfied. Taken together, these conditions then imply that $E[Y|\Delta_a,\Delta_b] = Pr(k_1+k_2+\nu(\Delta_a-\Delta_b)\geq\epsilon_0-\epsilon_1)$, which thus guarantees the existence of a function g such that $E[Y|\Delta_a,\Delta_b] = g(\Delta_a-\Delta_b)$.

[21] Except, of course, $\epsilon_0-\epsilon_1$.

This analysis demonstrates that, for an appropriately designed experiment, similar methods to those developed in our main text can be used to test whether a given set of choices can be rationalized by Kőszegi and Rabin's CPE concept. If the existence of a single-index representation $g(\Delta_a - \Delta_b)$ is rejected, that serves as a basis to reject Kőszegi and Rabin's notion of the reference point. Similar approaches may be used to generate means of testing unacclimating personal equilibrium or preferred personal equilibrium, although we omit the development of these approaches here due to their close similarity to the approach presented for CPE.

C Simulation Study of Power of Test

In this section, we assess the power of our test when applied to experiments like the one we ran. We simulate experimental data in which decisions are made according to a variety of parameterizations of standard prospect theory, conduct our test, and assess the rate of type-1 and type-2 error.

C.1 Parameters for Simulation

Simulating Gambles Presented and Reference Points: In our simulated experiments, simulated subjects face Δ-shifted versions of one of the 5 base scenarios presented in Table 1.

We randomly generate a true reference point governing the decision process (r^t) and a false candidate reference point that we wish to study (r^c). We randomly sample the values of Δ, r^t, and r^c from the distribution

$$\begin{pmatrix} \Delta \\ r^t \\ r^c \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 3.4 \\ 3.4 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \right) \tag{46}$$

The distribution matches the distribution used to simulate values of Δ and our two candidate reference points in the experiment that we deployed.

Simulating Choice: Given the randomly-generated gamble and reference points, we simulate choices as arising from a range of potential utility functions. We model utility of a given outcome in standard prospect-theory form:

$$\phi(x|r^t) = \begin{cases} (x - r^t)^\alpha & \text{if } x \geq r \\ \lambda(r^t - x)^\alpha & \text{if } x < r \end{cases} . \tag{47}$$

Choices are made to maximize expected utility, complete with a random-utility component: $\sum_{o \in O} p_o \cdot (\phi(c_o - r)) + \epsilon$. In this simulation, we do not include a direct-utility component. For our purposes, this is equivalent to assuming that local linearity holds exactly for direct consumption utility (ψ).

Given this specification of choices, the relevant parameters for simulations are λ, α, and the parameters governing the distribution of the additive error term ϵ. λ is sampled from the values 1, 1.5, 2, 2.5, and 3; this range is distributed around median estimates in the literature (approximately $\lambda = 2$, see Brown et al., 2020), and includes as one endpoint the case where loss aversion is not present ($\lambda = 1$). α is sampled from the values 0.6, 0.7, 0.8, 0.9, and 1, covering a range from relatively extreme diminishing sensitivity to none at all. Note that we do not sample situations where $\alpha = \lambda = 1$, because this results in linear reference-dependent utility in violation of Assumption 1.

Construction of the error term is made somewhat more complex by the fact that different values of α and λ change the scale of fixed utility differences. Thus, holding constant the mean and variance of ϵ, the rate of preference-reversals would not be held constant across different draws of α and λ. To address this issue, while simultaneously allowing for different within- and between-subject distributions of ϵ, we use the following approach to simulating these errors. For each set of simulated values, we first calculate the deterministic portion of the utility difference. Define parameter M to equal the standard deviation of that value, holding fixed all utility parameters and the baseline gamble, but varying the draw of Δ. For

each choice, we model the error process as being governed by the sum of two components:

$$\text{subject-specific shock: } \epsilon_i \sim N(0, (M_i \cdot s_i)^2) \tag{48}$$

$$\text{choice-specific shock: } \epsilon_j \sim N(0, (M_j \cdot s_j)^2) \tag{49}$$

Terms s_i and s_j are terms that are used to scale up or down the degree of variance in each component. When these values are set to 1, then the distribution of the error term is such that a 1-standard-deviation shock is scaled to a 1-standard-deviation difference in the deterministic utility component. s_i and s_j represent the scale of the degree of cross-subject and within-subject choice heterogeneity respectively. We consider three values of each scale term: 1, 2.5, and 4, resulting in nine combinations total.

Simulating Sample Size and Panel Length: Across simulations, we vary two features of the way datasets could be generated: the number of subjects included, and the number of questions posed to each subject. We consider potential numbers of subjects drawn from the values 50, 100, 300, 500, and 1,000, ranging from the size of comparatively small lab experiments to larger experiments only possible in online formats. We additionally consider a range of numbers of questions presented of 1, 2, 3, or 4.

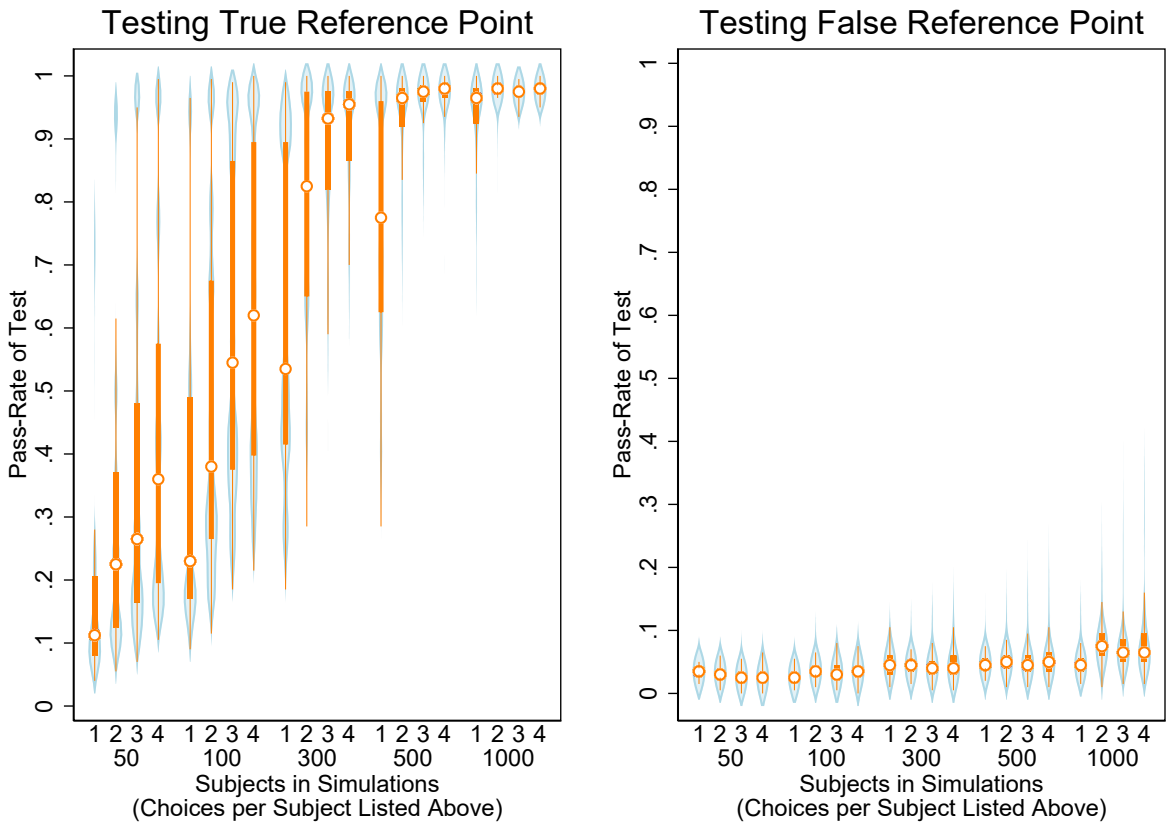
Summary of All Iterations: Across all dimensions varied above, there are 43,200 unique combinations possible: applying the correct or incorrect reference point (2) $\times$ 5 Baseline gambles $\times$ 24 combinations of α and λ $\times$ 9 versions of the error distributions $\times$ 5 potential sample sizes $\times$ 4 potential panel lengths. For each of the 43,200 combinations, we simulate 200 datasets for analysis, yielding a total of 8,640,000 simulated experiments. Within each batch of 200 datasets simulated under fixed parameters, we calculate an aggregate "pass rate" among those 200 applications of our test. An application is coded as passing if our test fails to reject the candidate reference point in our stage-1 test, but does reject the degenerate form of reference dependence screened in our stage-2 test.

C.2 Results of Simulations

Figure C.1 presents violin plots summarizing pass rates in our full set of simulations.

We begin by focusing attention on the left panel of the figure, which presents results for

Figure C.1: Assessing Pass-Rate of Test Across Simulations



Notes: The left panel presents results when the test is applied to the true reference point used in the simulation—i.e., cases where the test would ideally pass. The right panel presents results when the test is applied to a false candidate reference point that is statistically independent from the true reference point used in the simulation—i.e., cases where the test would ideally fail. Within each panel, for a range of the number of subjects and the number of observations per subject, we summarize the distribution of pass rates achieved in the 200 iterations run for each combination of potential simulation parameters. The orange dots present the median pass rate, the thick portion of the orange line represents the interquartile range, and the thin orange line extends to the upper- and lower-adjacent values. Behind each line is small kernel-density representation of the distribution.

the cases where we apply our test to the true reference point used in the simulations. In these situations, our test would ideally pass. This would fail to occur if our test generated a type-1 error by rejecting the true reference point, or if the first stage passed but the second stage generated a type-2 error by failing to reject the null of degenerate reference-dependence.

The x-axis of this figure covers the range of sample sizes considered, and varies both

the number of subjects in each simulation and the number of choices posed to each subject. Above each potential sample size, we summarize the distribution of pass rates across all sets of simulated parameter values. The orange dots present the median pass rate, the thick portion of the orange line represents the interquartile range, and the thin orange line extends to the upper- and lower-adjacent values. Behind each line is a small kernel-density representation of the distribution.

Summarizing this panel as a whole, we note that our pass-rate converges to a rate of approximately 95% relatively quickly as sample size increases. In our simulations with 500 or more subjects included, the pass rate is uniformly high regardless of the number of observations generated by each subject, and with 300 observations the pass rate is high as long as more than 2 observations are collected per subject. When only 50 or 100 subjects are included in the simulation pass rates are well below their ideal. This is largely influenced by being ill-powered to reject the null hypothesis in stage-2 of the test, a necessary step for counting an application as a "pass." This issue may conceivably be alleviated by collecting more observations per subject: we limited our simulation to only cases up to 4 observations per subject largely because the computation time slows down sufficiently quickly with the number of panel observations to make a high-iteration-simulation impractical, and not because we think collecting more data past this point is unproductive.

We next turn attention to the right panel of the figure, which presents results for the cases where we apply our test to an incorrect reference point simulated to be statistically independent from the true reference point. These simulations are somewhat more straightforward to characterize: across the sets of parameters and samples sizes considered, our ability to reject false reference points is uniformly high. Even with the smallest sample sizes considered, a false reference point is rejected in the vicinity of 95% of the time on median, with relatively little variation across the sets of parameters studied. While this degree of power is perhaps surprising, two simple forces contribute. First, as noted about, our stage-2 test has low power to reject a null of a constant choice probability, making acceptance of candidates rare for small sample sizes. Conceptually, while this makes "passing" true reference points harder at small sample sizes, it makes rejecting false reference points easier. Second, recall that our test can be understood to be asking "are all level-sets in $\Delta \times r$ space parallel lines of slope

1?"

We interpret these findings to suggest that, for reference-dependent utility functions of the type typically considered in this literature, the diagnostic value of our test for detecting the correct reference point is quite high even in relatively modest sample sizes.

D Supplementary Tables

This section contains supplemental tables referenced in text.

Table D.1: Summary Statistics from UAS Data

	(1)	(2)	(3)	(4)
	Survey Completion Status			Test for Differences
	Complete	Incomplete	All Recruits	
Basic Demographics				
Female	61.4	56.8	57.9	p = 0.14
Married	47.4	60.1	56.9	p = 0.00
Working	58.6	58.4	58.4	p = 0.96
US Citizen	97.0	97.9	97.7	p = 0.34
Hispanic or Latino	11.7	6.9	8.1	p = 0.01
Race				.
White Only	81.0	76.9	80.0	.
Black Only	8.9	12.6	9.8	.
Am. Indian or Alaska Native Only	1.2	0.9	1.1	p = 0.37
Asian Only	3.2	2.7	3.1	.
Hawaiian/Pacific Islander Only	0.5	0.9	0.6	.
Multiple Races Indicated	5.2	5.7	5.3	.
Education				.
< 12th grade	4.7	4.5	4.7	.
High school grad.	18.6	16.8	18.2	.
Some college	22.5	21.9	22.4	p = 0.08
Assoc. degree	14.8	15.6	15.0	.
Bachelor's degree	22.2	29.1	23.9	.
Master's degree +	17.2	12.0	15.9	.
Household Income				.
< $10,000	5.9	9.3	6.8	.
$10,000 - $24,999	13.0	15.9	13.7	.
$25,000 - $49,999	20.6	21.0	20.7	p = 0.14
$50,000 - $74,999	21.0	20.1	20.8	.
$75,000 - $99,999	14.0	12.6	13.7	.
$100,000 +	25.2	20.7	24.1	.
Age				.
18-29	7.3	15.0	9.2	.
30-39	16.6	22.2	18.0	.
40-49	17.4	16.5	17.2	p = 0.00
50-59	21.9	18.3	21.0	.
60+	36.8	27.9	34.6	.

Notes: This table presents summary statistics characterizing the demographic features of our sample. With the exception of p-values, all numbers presented are the percentage of respondents with a given row's classification. The first panel characterizes a series of binary demographic variables, and the panels that follow present tabulations of individual categorical variables. The first column presents results for subjects included in our primary analyses. To help assess selection into our study, the second and third columns present results for the subjects who were contacted but did not complete the study and all contacted subjects, respectively. The final column provides p-values for Fisher Exact Test for differences in the distribution of the categorical variable by survey completion status.

Table D.2: Base-Scenario-Specific Test Results

		Stage-specific p-value		
	Scenario Number	Stage 1	Stage 2	Passed?
Treatment Arm: Goal Salient				
	1	0.5164	0	Yes
	2	0.1986	0	Yes
Ref Pt: Goal	3	0.6284	0	Yes
	4	0.8878	0	Yes
	5	0.0403	0	No
	Pooled	**0.7778**	**0**	**Yes**
	1	0.1784	0.7024	No
	2	0.6404	0.8658	No
Ref Pt: Average Earnings	3	0.5201	0.9414	No
	4	0.9594	0.4757	No
	5	0.1103	0.2702	No
	Pooled	**0.5057**	**0.5157**	**No**
Treatment Arm: Average Earnings Salient				
	1	0.9118	0.6994	No
	2	0.254	0.2737	No
Ref Pt: Goal	3	0.4573	0.3051	No
	4	0.6431	0.9938	No
	5	0.1986	0.494	No
	Pooled	**0.4911**	**0.6852**	**No**
	1	0.0602	0.7268	No
	2	0.992	0.7059	No
Ref Pt: Average Earnings	3	0.882	0.473	No
	4	0.7511	0.5861	No
	5	0.7102	0.2785	No
	Pooled	**0.7024**	**0.1568**	**No**
Treatment Arm: No Reference Point Salient				
	1	0.8467	0.0461	Yes
	2	0.3018	0.268	No
Ref Pt: Goal	3	0.8479	0.1777	No
	4	0.0365	0.7569	No
	5	0.1969	0.6592	No
	Pooled	**0.2634**	**0.3588**	**No**
	1	0.2186	0.2896	No
	2	0.232	0.904	No
Ref Pt: Average Earnings	3	0.8431	0.8728	No
	4	0.5201	0.7461	No
	5	0.1258	0.4217	No
	Pooled	**0.1074**	**0.4602**	**No**

Notes: This table presents the the results from our proposed test across different treatment arms and candidate reference points. The first panel present results from the treatment arm where the goal is made salient. The second panel present results from the treatment arm where average earnings are made salient. The third panel presents results from the control arm where neither reference point is made salient. Within each panel, we separately present results for each of the five base scenarios. The first column indexes the base scenario group, the second and third columns present p-values from the first and second stage of our test, respectively, and the final column indicates whether the test as a whole has passed.