An Approach to Testing Reference Points

Alex Rees-Jones and Ao Wang*

April 27, 2022

Abstract

The application of reference-dependent models is often complicated by the modeler's uncertainty regarding the reference point (referent) that agents adopt. We develop a powerful and minimally parametric approach to testing whether decisions could be rationalized by a general reference-dependent model with a specific referent. Our approach builds from the observation that, when both payoffs and the true referent are randomly varied, a marginal increase in all payoffs will have an equivalent effect as a marginal decrease in the referent. The observation that this equivalence holds at all payoff/referent combinations, when applied to decisions over properly constructed gambles, allows us to generate our test through modifications to existing tools for rejecting single-index representations. We assess the performance of this test in a simulation study and find that it is highly diagnostic even in the comparatively small sample sizes that are common in experimental economics. We then utilize this approach in an online experiment in which we randomly vary the salience of both goal-based and expectations-based referents. In this experiment, we confirm the common assumption that salient goals could serve as reference points. Illustrating the importance of salience, we reject that either reference point is adopted when it is not salient. Perhaps surprisingly, we reject the adoption of expectations as a reference point even when they are salient.

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

*Rees-Jones: University of Pennsylvania and NBER, email: alre@wharton.upenn.edu. Wang: University of California, Berkeley, email: ao.wang@berkeley.edu. We are grateful to Ori Heffetz, Ted O'Donoghue, and especially Francesca Molinari for helpful guidance in 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 USC, the UAS, or any other entity.

Models of reference-dependent behavior, most notably prospect theory, are pervasive in modern behavioral economics. The defining feature of these models is that the input to utility is an amount compared to another value, normally called a reference point. In many settings, this relative treatment of inputs to utility provides an explanation for behavior unexplainable by more standard expected-utility approaches (Camerer, 2004). A large amount of experimental-economic research has assessed these models and found strong support for their inherent psychological processes (Barberis, 2013). This evidence suggests that prospect theory has significant potential to be predictive in a wide variety of economic applications.

Despite this foundational support, reference-dependent models of this type are relatively infrequently employed in field settings with large-scale microeconomic datasets. This is undoubtedly influenced by the difficulty associated with credibly establishing a reference point. Existing evidence supports the adoption of different referents in different situations, and in many potential applications it is theoretically unclear which referent is at play.[1] Despite the substantial efforts of behavioral economists, a complete theory of reference point formation is not yet available. Until this uncertainty about the nature of the reference point is resolved, the practical value of models that rely critically on knowledge of the reference point is limited.

In this paper, we propose a methodology that facilitates the development of such theories of reference point formation. Our method 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 is minimally reliant on parametric assumptions and extremely robust to heterogeneity. In short, out test assesses whether the most fundamental predictions arising from the adoption of a candidate reference point are sufficiently violated to reject the candidate.

In section 1, we describe the intuition behind our test. In our view, the defining charac-

[1] Initial approaches to modeling reference dependence typically treated status-quos or endowments as reference points (see, e.g., Kahneman & Tversky, 1979; Kahneman et al., 1990; Kahneman & Tversky, 1992). Subsequent work has developed a variety of other alternatives, including goals (Heath et al., 1999; Hsiaw, 2013; Allen et al., 2017; Markle et al., 2018; Hsiaw, 2018) and expectations (Kőszegi & Rabin, 2006, 2007, 2009; Marzilli Ericson & Fuster, 2011; Crawford & Meng, 2011; Pagel, 2017, 2018). In practice, the choice between these and other options is often difficult and subjective. See Brown et al. (2020) for a thorough summary of the reference points used across empirical applications.

teristic 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 takes the form $u(c-r)$.[2] While many approaches to tests of reference points rely on the functional form of u—for example, relying on the assumption that the first or second derivatives change discontinuously at zero due to loss aversion or diminishing sensitivity, respectively—our goal is to be agnostic about the functional form beyond basic regularity conditions.

Even within this minimally parametric framework, 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 prediction could be easily examined in cases where utility is thought to be observed (e.g., when testing for reference dependence in life-satisfaction or happiness data), but our focus will be on the more standard economic framework where latent random utility is assumed to rationalize choices. Within such a framework, similar strong restrictions on the level-sets of choice functions arise when subjects are presented with appropriately constructed choices. These restrictions hold essentially without modification even in the presence of rich heterogeneity in individual utility parameters. This suggests an immediate strategy for robust inference regarding candidate reference points: randomly varying consumption and the candidate reference point, and then rejecting the reference point if this property is statistically rejected.

In Section 2, we describe the formal statistical framework we adopt in order to test these predictions. The key econometric insight arising from the framework in section 1 is that, if reference-dependent decision-makers are presented choices between gambles where all consumption values are perturbed by a common value Δ, then choice probabilities will admit a single-index representation with the level set properties just discussed (in $\Delta \times r$ space). Based on this insight, we can proceed by modifying existing, powerful non-parametric

[2]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, the formal approach presented in section 2 accommodates utility influenced by absolute and relative components.

techniques for formally testing this structure. Conceptually, the approach may be understood as estimating the single-index function using kernel methods and examining whether the fit achieved is sufficiently statistically unlikely to reject the null hypothesis of a correctly specified model. We adopt the work of Fan & Li (1996) to derive an analytical formula for the associated p-value, modified to accommodate additional structure imposed by our model and the clustering issues that arise in our domain.

In Section 3, we provide results from a simulation study of the performance of our estimator. Despite the common intuition that structural approaches relying on non-parametric methods are too demanding of data to be broadly used in lab experiments, we document that favorable rates of type-1 and type-2 error are achieved with sample sizes common in modern online experiments. Intuitively, this is because under the null hypothesis of correct reference-point specification, the function that is non-parametrically specified is univariate even as the sample size grows large. The curse of dimensionality is therefore avoided.

Given these encouraging simulation results, we designed and deployed an experiment as a field-test of this estimator. We describe the design of this experiment in Section 4. In this experiment, subjects chose between a sure option and a 50-50 gamble with all payoffs constructed in the manner dictated in Section 1. Subjects are also presented with randomized values of two variables that have been used as reference points in prior literature: goals and expectations. Based on our reading of existing literature, we believed each candidate could be adopted as reference points in at least some situations. 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 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 introductory materials.

In Section 5, we document our results. We deployed this experiment to 1,001 subjects in the Understanding America Study, a panel survey that aims to provide a venue for deploying studies like ours to representative samples of Americans.

Using these data and our econometric approach, we confirm a claim that has been presented in prior papers: that goals can serve as reference points. When goals are made salient, our approach fails to reject the hypothesis that choices are rationalized with a reference-dependent model with the randomly-varied goals used as the reference point. Emphasizing the importance of salience, however, our approach strongly rejects that hypothesis when goals are not salient (which includes cases where an alternative reference point is made salient or where no option is salient), establishing that non-salient goals are not reference points. 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.

Perhaps more surprising results arise when examining tests of expectations as reference points. While we do find some degree of reaction to expectations when they are salient, we reject the null hypothesis that choices are rationalized with a reference-dependent model with the randomly-varied expectations used as the reference point regardless of salience condition. These results raise questions about the use of expectations-manipulations like ours as reference points in the existing literature.

In Section 6, 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 the general 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: a large and successful technical literature has been developed on the formal non-parametric estimation of single-index models (see, e.g., Ichimura, 1993; Fan & Li, 1996; Horowitz, 2001; Horowitz & Mammen, 2004, 2010), but despite this success few field applications of these techniques have been found. As we document, with some modification this literature can be used to develop a broadly portable and easily implementable testing framework for one of the most core questions in behavioral economics.

1 Intuitive Explanation of Approach to Identification

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

Our goal is to understand the key sources of identifying power when examining a model that is reference dependent. We adopt a specific and reasonably broad definition of what it fundamentally means to be reference dependent: we model reference dependence as meaning that the input to our function of interest is relative rather than absolute. Beyond technical assumptions, the core substantive assumption we wish to make is that the observed utility input C and utility itself Y satisfy $y = \phi(c - r)$, where R is a reference point and ϕ is a monotone function. Throughout this discussion upper case letters denote random variables and lower case letters denote their specific realizations.

1.1 Intuition in Case with Direct Observation of Reference-Dependent Process

Most economic applications treat utility as fundamentally unobservable, and thus assume that Y 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 Y is directly observed.

In such a case, the consequences of correct specification of the reference point can be completely characterized by their implications on properly defined *level sets* of the variable Y.

Definition 1. *Define the level set of Y, evaluated for the specific value y and over hypoth-*

esized reference point R, to be the set of all (c, r)-tuples satisfying $y = \phi(c - r)$. Formally, denote this as $\mathcal{L}^Y_R(y) = \{(c, r) \in c \times r : y = \phi(c - r)\}$.

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

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

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

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 $y = \phi(c - r)$ for *some* reference point, then the reference point considered must be the wrong one.

1.2 Intuition in Case with Reference-Dependent Process Governing Choice of Gambles

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 in when latent and more sophsiticated 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 outcomes. Each gamble is defined by the set of possible outcomes $\mathcal{G} = (p_o, c_o)_{o\in O}$, denoting the probability (p_o) and the consumption (c_o) 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. *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*

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

ψ captures a standard direct utility function. ϕ captures a reference-dependent utility function. ϵ is the realization of an i.i.d random variable as in standard random utility models.

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.

Given a shifting parameter $\Delta \in \mathbb{R}$ and a base gamble $\mathcal{G}$, we denote 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

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

Note that the reference-dependent components can be consolidated into a single-index function (ν), allowing this equation to be expressed as

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

Note that this structure remains more complicated than the single-index representation derived in Section 1.1. However, consider an additional assumption that is commonly assumed to hold:

Assumption 2. *Over the support of* Δ, $\sum_{o \in O^{\mathcal{G}_1}} p_o^{\mathcal{G}_1} \cdot (\psi(c_o^{\mathcal{G}_1} + \Delta) - \sum_{o \in O^{\mathcal{G}_0}} p_o^{\mathcal{G}_0} \cdot (\psi(c_o^{\mathcal{G}_0} + \Delta) \approx k$ *for an arbitrary 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 approximately equal. This guaranteed to hold exactly if consumption utility (ψ) is linear. Less restrictively, it holds when consumption utility (ψ) is 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 property holds in common economic models. Concretely, arguments like that of the Rabin Calibration Theorem (Rabin, 2000) suggest that changes to the curvature of the utility function are negligible if baseline consumption is shifted by several dollars.

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

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

In short, the same single-index structure that arose over (c, r) in the direct utility case arises over (Δ, r) in the latent utility case. This yields the key finding underlying our approach:

Proposition 1. *Consider choices between gambles $S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$. Assume that choices are governed by utility satisfying Assumptions 1 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. There exists a single-index function g such that $E[Y|\Delta, r] \approx g(\Delta - r)$.*

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. Following the strategy above, our simulated individual will face choices between Δ-shifted versions of these gambles ($S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$, respectively). 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 relatively standard prospect-theory value function featuring loss aversion, diminishing sensitivity, and no direct-utility component:

$$\phi(c|r) = \begin{cases} (c-r)^{0.6} & \text{if } c \geq r \\ -2(r-c)^{0.6} & \text{if } c < r \end{cases} . \tag{7}$$

As above, assume the individual chooses the risky gamble only if $U(\mathcal{G}_1|r) \geq U(\mathcal{G}_0|r)$, 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 normally distributed with a mean of 0 and a standard deviation of 4.

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 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. To understand the shape of this function, imagine that the reference point is fixed at 0. For draws of Δ below -2, all outcomes of both the safe and risky option are coded as losses. Within this region, the assumed diminishing sensitivity of the value function results in risk-loving behavior. This leads to a higher chance of choosing the risky option, particularly when considering gambles over relatively small losses when the individual is most risk loving. For values of Δ between -2 and 1, the two outcomes in the risky option fall on either side of the reference point. The kink at zero in the utility function results in first-order risk aversion in this region, causing the precipitous decline in the probability of choosing the risky option observed. When Δ exceeds 1, all outcomes are in the gain domain. The first-order risk aversion around the reference point is no longer relevant, and choices are now primarily influenced by the standard second-order risk aversion that occurs over gains. While standard assumptions on prospect theory lead to this understandable qualitative structure, the precise shape of this function will be determined by the exact parametric assumptions made on the utility function. This motivates our desire not to use this function directly for identification, given the uncertainty that exists about true parametric form.

The right panel of Figure 2 plots the level sets associated with each dot in $\Delta \times r$ space. 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, with darker colors denoting higher probability of choosing the risky option. As in the observed-utility case, this results in a pattern of parallel lines of slope 1. Importantly for robust identification, this precise pattern would remain even as parametric assumptions about, e.g., diminishing sensitivity or loss aversion were varied—such changes would merely change the utility values of different level sets, but not the level sets themselves. This motivates our interest in using this pattern for identification; in contrast to approaches based on examining the the choice probability function directly, this approach does not require the researcher to impose detailed functional form assumptions.

1.3 Robustnesses to Heterogeneity in Auxiliary Utility Parameters

In addition to providing the stark reduced-form predictions discussed in the prior section, the representation guaranteed by Proposition 1 has highly desirable properties for its ability to accommodate heterogeneity.

Proposition 2. *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 and 2, 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 single-index function* g *such that* $E[Y|\Delta, r] \approx g(\Delta - r)$*.*

Proof. For each $\theta \in \Theta$, Proposition 1 guarantees the existence of an g_θ such that $E[Y|\Delta, r, \theta] \approx g_\theta(\Delta - r)$. Applying the law of iterated expectations, notice that $E[Y|\Delta, r] = E[E[Y|\Delta, r, \theta]|\Delta, r] \approx E[g_\theta(\Delta - r)|\Delta, r]$. Using the assumption that θ is statistically independent from Δ and r, we may now define a new single-index function $\bar{g}(\Delta - r) = E[g_\theta(\Delta - r)|\Delta, r] \approx E[Y|\Delta, r]$. □

The reasoning above immediately implies that the approach suggested by Proposition 1 remains valid for a population of agents who all satisfy Assumptions 1 and 2, 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 (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.

This result also demonstrates a key way in which we are relying on experimental variation in our design of this test. Our approach is predicated on the idea that Δ and r are under the control of the experimenter, and thus that they are capable of being experimentally assigned in a manner that is statistically independent from individual heterogeneity. If this were not the case, then the single-index structure could be broken. Conceptually, we are focusing on the prediction that all values of Δ and r that satisfy $\Delta - r = c$ should be viewed equivalently and lead to the same distribution of responses. If, for example, a Δ of 1 were only assigned to risk-tolerant individuals and a Δ of 2 were only assigned to risk-loving individuals, there would be no reason to expect that the tuples $(\Delta = 1, r = 1)$ and $(\Delta = 2, r = 2)$ would be associated with the same behavior in population, even though they would be viewed equivalently within-type. This demonstrates a fundamental challenge in attempting to modify these techniques for use in observational data, while also emphasizing a fundamental strength of the approach in an experimental setting.

2 Proposed Estimation Strategy

In this section, we present our formal estimation strategy. This strategy is based on the intuition expounded in the prior section. The key element of this intuition is that, in our broad class of models, the data generating process (DGP) admits a single-index representation if the reference point has been correctly specified. This observation allows us to link the empirical and experimental literature testing reference-points to the econometric theory literature on specification testing of single-index models. This literature provides a means for formally testing our null hypothesis of interest: that the DGP admits a representation with the required structure $E[Y|c,r] = g(c-r)$. As was demonstrated in the previous section, in the latent utility case that will be our main focus, we may simply replace the term c with Δ and otherwise proceed identically.

Our econometric approach builds heavily 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 $Y = g_1(x_1, x_2)$ not imposing single-index structure and comparing that to a kernel-smoothed approximation of $Y = g_2(X\beta)$ that imposes the single index structure. To modify this to our setting, we simply take advantage of the additional restrictions imposed by a reference-dependent model: not only should the DGP admit a representation as $Y = g_2(X\beta)$, but furthermore the linear component is specifically $X\beta = x_1 - x_2$. 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.

Full details of the derivation of our test statistic and proof of its asymptotic distribution are available in the Appendix.

The goal of this analysis is to assess the value of a finite-sample estimate of

$$E[v \cdot f(c-r)E[v \cdot f(c-r)|c,r]] \tag{8}$$

where $f(c-r)$ is the p.d.f. of (c-r) and v is the approximation error induced by assuming this structure ($v = E[y|c,r] - E[y|c-r]$). Note that if the DGP admits a representation of $Y = g(c-r)$, v is zero for any c and r. Consequently, this product will also be zero. Given a kernel-based approximation to $g(c-r)$, approximation error will lead to this product not being identically zero, but instead distributed around zero. In contrast, if the DGP does not admit such a representation, this product is positive and growing with sample size. Our test proceeds by generating a test statistic that has a known distribution around zero under the null hypothesis, so we may establish how unlikely the fit is under that null hypothesis.

To begin, define

$$\widehat{E}[Y_{(i,j)}|c_{(i,j)} - r_i] = \frac{[(N-m)a]^{-1}\sum_{i'\neq i}\sum_{j'} Y_{i',j'}K^{\alpha}_{(i,j),(i',j')}}{\widehat{f}_{\alpha}(c_{i,j} - r_i)}. \tag{9}$$

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 $N = n \cdot m$. Additionally define the kernel-density estimate of f_α as

$$\widehat{f}_{\alpha}(c_{(i,j)} - r_i) = \frac{1}{(N-m)a}\sum_{i'\neq i}\sum_{j'} K^{\alpha}_{(i,j),(i',j')} \tag{10}$$

in which k^α is the univariate guassian kernel and $K^{\alpha}_{(i,j),(i',j')} = k^{\alpha}(\frac{(c_{(i,j)}-r_i)-(c_{(i',j')}-r_{i'})}{a})$. Denote the bandwidth used for the kernel regression assuming single-index structure by a and the bandwidth used for the two-dimensional density estimate by h.

Given these definitions, we may now generate an estimate of $E[vf_\alpha(c-r)E[vf_\alpha(c-r)|c,r]]$ with

$$I = Nh\frac{1}{N(N-m)h^2}\sum_{i}\sum_{i'\neq i}\sum_{j}\sum_{j'}[\bar{v}_{(i,j)}\widehat{f}_{\alpha}(c_{(i,j)} - r_i)][\bar{v}_{(i',j')}\widehat{f}_{\alpha}(c_{(i',j')} - r_{i'})]K_{(i,j),(i',j')} \tag{11}$$

where $\bar{v}_{(i,j)} = Y_{(i,j)} - \widehat{E}[Y_{(i,j)}|c_{(i,j)} - r_i]$. $K_{(i,j),(i',j')}$ is a product of two univariate normal kernels where the bandwidth is h. Under our null hypothesis that the DGP satisfies $E[Y|c,r] = g(c-r)$, this statistic is asymptotically normally distributed with mean zero and a standard deviation of $\sqrt{2 \cdot (\widehat{\sigma_a} + \widehat{\rho_a^2})}$, where

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

and

$$\widehat{\rho_a^2} = \frac{(m^2-1)h}{N(N-m)(m-1)^2 h^3} \sum_i \sum_{j_1 \neq j_2} \bar{v}_{(i,j_1)} \widehat{f}_\alpha(c_{(i,j_1)} - r_i) \bar{v}_{(i,j_2)} \widehat{f}_\alpha(c_{i,j_2} - r_i) \sum_{i' \neq i} \sum_{j_1' \neq j_2'}$$
$$\bar{v}_{(i',j_1')} \widehat{f}_\alpha(c_{(i',j_1')} - r_i') \bar{v}_{(i',j_2')} \widehat{f}_\alpha(c_{(i',j_2')} - r_i') K\Big(\frac{c_{(i',j_1')} - c_{(i,j_1)}}{h}, \frac{c_{(i',j_2')} - c_{(i,j_2)}}{h}, \frac{r_i' - r_i}{h}\Big) \int k^2(s_2) ds_2$$

In the latter equation, $k(.)$ is the univariate Gaussian kernel.

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. DGP. For experimental applications that elicit multiple observations per subject, i.i.d. is 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 cluster.

2.1 Interpretation of Results of the Test

Upon calculating the test statistic described above, one of two outcomes may emerge.

One possibility is a rejection of the null hypothesis of the single-index representation. This outcome reveals that the data are comparatively unlikely to be observed under the joint null hypothesis of our flexible reference-dependent model and the specific suggested reference point. Because our proposed reference-dependent model is so general, we will typically discuss this as a rejection of the latter element of the null hypothesis: that the reference point was correctly specified.

The other possibility is a failure to reject the null hypothesis of a single-index representation. This generally means that one cannot reject the possibility that the candidate reference point was used. Two considerations are needed for interpreting the importance of that claim. First, as with any null result, one must consider the power of the test to reject false null hypotheses. We will provide more analysis relevant to assessing power in Section 3. Second, more specific to our model, we must rule out a special class of models which are technically "reference dependent" as we have defined the term, but trivially so. Recall that we consider a model reference dependent if it takes the form $Y = g(c - r)$ for some function f. Note that under this definition, a DGP in which Y is constant is considered reference dependent. For example, in our latent utility model, if local linearity is satisfied and the reference dependent element of utility (ϕ) is set to zero, the probability of choosing a given gamble is fixed. Failing to reject this trivial form of reference dependence should not be taken as evidence in favor of the concept of reference dependence as discussed in the behavioral economics literature.

To deal with this issue, we propose a second stage of our estimation process to be undertaken upon a failure to reject the null. In this second stage, we formally test if we can reject the null that $g(c-r)$ is a constant function. Conceptually, this test uses the same approach as the first-stage. If we define u to be the approximation error induced by fitting a constant function ($u = E[y|c-r] - E[y]$), the equation $E[uf(c-r)E[u|c-r]]$ is identically zero if the true DGP is constant. In contrast, if the DGP is not constant, this equation is positive and growing in N. Exactly analogous methods drawn from Fan and Li (1996) can be applied to characterize the distribution of a finite-sample estimate of $E[uf(c-r)E[u|c-r]]$ given a kernel-smoothed approximation to $E[u|c-r]$. Formally,

$$I_\mu = \frac{1}{N(N-m)a} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} (Y_{(i,j)} - \hat{\mu})(Y_{(i',j')} - \hat{\mu}) K(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \tag{12}$$

where $\hat{\mu} = \frac{1}{N} \sum_i \sum_j Y_{(i,j)}$, $X_{(i,j)} = c_{(i,j)} - r_i$, $K(.)$ is univariate Gaussian kernel, a is bandwidth. Under the null that DGP is a constant, $N\sqrt{a} I_\mu$ is normally distributed with zero mean. The variance of the normal distribution is estimated using the expression

$2(\hat{\sigma}_\mu^2 + (m^2-1)a\hat{\rho}_\mu^2)$, where

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

and

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

In situations where our test is well powered, we view the simultaneous failure to reject the first-stage test and a rejection of the second-stage test as evidence supporting the idea that a non-trivial reference-dependent model with the specified reference point can provide a good explanation of the data.

3 Simulation Study of Power of Test

In this section, we provide a careful examination of the power of this test in a set of gambles that could potentially be presented to subjects. These gambles are those that we presented in our experiment described in the next section.

3.1 Parameters for Simulation

Simulating Gambles Presented and Reference Points: We create 5 baseline binary-choice scenarios, each presenting a choice between some amount that will be earned with certainty and a 50-50 gamble between a comparatively high and low amount. We will denote the baseline payoff of the sure gamble as q_a, and the payoffs arising from the two states of the 50-50 gamble as q_b and q_c. The values simulated for these parameters across the 5 scenarios are presented in Table 1.

As described in the previous sections, it is key to our estimation strategy that choices be made over a series of Δ-shifted gambles. When simulating a specific gamble, we randomly draw a value of Δ and create the two gambles ($S(\Delta|\mathcal{G}_0)$ and $S(\Delta|\mathcal{G}_1)$) by taking two gambles for each scenario in Table 1 as the baseline gambles ($\mathcal{G}_0$ and $\mathcal{G}_1$).

In addition to specifying the gambles, we additionally randomly generate a true reference point governing the decision process (r^t) and a candidate reference point that we wish to study (r^c). As a benchmark, the candidate reference point is independent of the true reference point. We randomly sample the values of Δ, r^t, and r^c from the distribution

$$\begin{pmatrix} \Delta \\ r^t \\ r^c \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 3.4 \\ 3.4 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \right) \tag{13}$$

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:

$$\phi(x|r^t) = \begin{cases} (x - r^t)^\alpha & \text{if } x \geq r \\ \lambda(r^t - x)^\alpha & \text{if } x < r \end{cases} . \tag{14}$$

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.

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:

$$\text{subject-specific shock: } \epsilon_1 \sim N(0, (M \cdot s_1)^2) \tag{15}$$

$$\text{choice-specific shock: } \epsilon_2 \sim N(0, (M \cdot s_2)^2) \tag{16}$$

Terms s_1 and s_2 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_1 and s_2 represent the scale of the degree of cross-subject and within-subject choice heterogeneity respectively. We consider four combinations of values for these parameters: $(s_1, s_2) = (1, 2), (1, 1), (\frac{1}{2}, 1)$, or $(\frac{1}{2}, \frac{1}{2})$. Across these four values we vary both the overall variance of the error distribution, as well as whether subject- and choice-specific shocks are of comparable variance or whether choice-specific shocks show greater variance (as they do in our online experiment).

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 19,200 unique combinations possible: applying the correct or incorrect reference point (2) $\times$ 5 Baseline gambles $\times$ 24 combinations of α and λ $\times$ 4 versions of the error distribution $\times$ 5 potential sample sizes $\times$ 4 potential panel lengths. For each of the 19,200 combinations, we simulate 200 datasets for analysis, yielding a total of 3,840,000 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.

3.2 Results of Simulations

Figure 3 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 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 300 or more subjects included, the pass rate is uniformly high regardless of the number of observations generated by each 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." Despite this issue, note that even in these small samples the pass rate generally exceeds 90% when 4 observations are generated per subject. In sum, across a range of parameters spanning common applications of prospect theory, our full test achieves a rate of type-1 error (failing to pass a true reference point) in the vicinity of 5% in all but the smallest sample sizes.

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 more than 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?" For a candidate referent that is statistically independent from the true referent, the relevant level sets will have slopes of zero, reflecting the fact that subjects do not respond to this reference point. The high ability of our test to reject candidate referents partially derives from the fact that relatively few observations are needed to tell the difference between a slope of 1 and a slope of 0.

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.

4 Applying Our Approach in an Online Experiment

In this section, we describe an online experiment that we designed and deployed to serve as a testing ground for our approach. This serves as a demonstration of how to run an experiment optimized for this econometric technique, and additionally reveals new insights into reference point adoption.

4.1 Experimental Design

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 both heads and tails mapped to the same amount of money. Figure 4 shows a screenshot of the initial explanation of this format.

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"—a coinflip would be simulated and they would receive a bonus corresponding to the gamble that they chose. Payment for taking the full study consisted of a $4 fixed payment plus this bonus.

After the presentation of these initial instructions, subjects faced a series of three questions meant to verify their understanding of the decision 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 $X 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 $Y.*

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. This salience is achieved by including large red text over the gamble choice reminding the subject of either their goal or the average. In those conditions, one potential reference point is entirely ignored after this page and the other has a constant vivid reminder 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 groups of four questions. Within each group of 4 questions, all gambles are Δ-shifted gambles based on a random draw of Δ. The values of Δ and both reference points were randomly generated according the distribution previously applied in our simulation (see equation 13). The base gambles used vary by group and are presented in Table 1.

The theory of Section 2 details how to apply our test to data in which all gambles are Δ-shifted from a single base gamble. We chose to generate data with multiple groups of questions, each Δ-shifted from a different base gamble, for two primary reasons. First, by doing this we generate five separate opportunities to apply our test in a single experiment. As long as sufficient power is obtained with relatively few observations per subject (as we verified in Section 3), we believed it was worthwhile to sacrifice the power benefits of devoting all questions to a single test in order to generate the opportunity to study the performance of our test in the context of multiple base gambles. Second, by mixing together Δ-shifted values from multiple base gambles and randomizing question order, we obfuscated the fact that amounts were not randomly generated in an unrestricted way. We believe that many subjects who saw 20 Δ-shifts of a common base gamble in sequence would come to understand how the randomization was occurring, which could conceivably influence behavior. In contrast, we do not believe that subjects in our experiment would be able to make similar inference.

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.[3]

4.2 Experimental Deployment

In December 2020 and January 2021, we deployed our experiment in the Understanding America Study (UAS), an online panel of American Households.[4] 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, with periodic reminders provided. The study was closed shortly after the target sample size was attained, ultimately resulting in 1,001 complete observations and a 75% response rate.

Table 2 summarizes basic demographics of our respondents. As is seen across panels of this table, our sample is demographically diverse. Relative to the full U.S. population, participants in our survey are notably 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.

Prior to deployment, our study was preregistered on aspredicted.org.[5] This preregistration specified our sample size, precise analyses of interest, and default values for the tuning parameters in our non-parametric approach.

4.3 Experimental Results

Using our experimental data, we conduct a series of hypothesis tests that are informative about reference point adoption. 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 performance 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 existing

[3] Available at https://uasdata.usc.edu/survey/UAS+287.
[4] For a detailed description of the UAS, see Alattar et al. (2018).
[5] Available at https://aspredicted.org/7pc6i.pdf.

literature, we expected that either could be adopted as a reference point under the right conditions. However, we also expected that they would only be adopted when they are made sufficiently salient. When neither referent is made salient we expected that neither would be adopted.

To test these hypotheses, we apply our approach to test for the adoption of each reference point separately across each treatment arm, and among each of the five question groups within the treatment arm (with a question group consisting of the four Δ-shifted questions corresponding to the same base gamble).

Table 3 presents the results. In the top panel, we present results for the control arm. Column 2 shows the p-values associated with stage-1 of our test, in which a rejection of the null hypothesis means a rejection of the hypothesis that choices can be represented with a functional of form $g(\Delta - r)$. Across all questions groups and for both potential reference points, this test is only rejected for a single question group when the goal reference point is applied. In general, this pattern of results leaves open the possibility of adoption of both reference points. However, examining column 3, this possibility is decisively ruled out. For 9 of the 10 tests, we fail to reject the null of the degenerate form of reference dependence that the second-stage is meant to detect: complete unresponsivity to $(\Delta - r)$. As a result, for 9 of these 10 tests, the final implication is that the composite test did not pass (as indicated in column 4), meaning that the test does not suggest that the reference point could be adopted. These results are broadly in line with our expectation that reference dependence relative to our randomly generated referent will not arise when referents are not made salient.

Turning next to the panel presenting results for the arm where goals are made salient, a different pattern emerges. When our test is applied to the goal reference points, the test passes for four of the five question groups, supporting the possibility that our randomly assigned goals are adopted as reference points when they are salient. In contrast, for all five of the questions groups, we can reject the hypothesis that the average reference point is adopted when the goal reference point is made salient. This pattern of results is in general supportive of the idea that goals can be adopted as reference points under the right circumstances.

Turning finally to the panel presenting results for the arm where averages are made salient, a somewhat surprising pattern emerges. In contrast to our results for the goal arm, in this arm no tests pass. We may reject that choices were made in a reference-dependent manner relative to either reference point—the non-salient goal reference point or the salient average reference point.

4.3.1 Direct Examination of Level Sets

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." To help shine light in the black box, we now document the fundamental features of the data that drive the statistical rejections of reference points documented in Table 3. To do so, we directly examine contour plots of choice probabilities mapped over $\Delta \times r$ space to search for the parallel level-sets of slope 1 that were key identifying feature highlighted in our intuitive description of our approach.

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 measures of Δ and r. Figure 5 presents these estimated choice probabilities plotted over a fine grid.[6] 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. In all cases, we pool all 5 questions groups when conducting these estimations.

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. Indeed, 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 inability of our statistical test to reject the goal reference point when it is salient can be understood to derive directly from this pattern: the empirical relationship is "close enough" to the theoretical prediction

[6] Grid increments are 0.05 in all dimensions.

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, here there are no obvious parallel lines of note. The figure as a whole does not resemble the theoretical predictions emphasized in Figure 2, and indeed it is “different enough” that the difference can be statistically detected. When goals are not made salient, this is the feature of our data fundamentally driving their rejection as candidate reference points 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. The key observation to note from these two panels is that both bear little resemblance to the predicted structure under a correctly specified reference point. We see no suggestion of the pattern of parallel lines of slope 1, and as in the last case considered the empirical patterns are “different enough” that they drive the rejection of this reference point in our formal statistical tests.

5 Discussion

Reference dependence is among the most well-trodden 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, allowing for principled hypothesis testing of proposed candidates. Applying this technique in an online experiment, we found clear support for salient goals serving as reference points. We additionally rejected that even salient averages serve as reference points in our setting. We emphasize that a finding that expectations do not serve as reference points our setting does not imply that they can never serve as reference points in other settings. We additionally emphasize that some experiments present or conceptualize averages in different manners than we have, and that the leading models of expectations-based reference points treat them as endogenously determined and not exogenously manipulable. These caveats aside, we believe that our finding is surprising in light of some of the existing literature,[7] and further emphasizes the need for further research that allows to best predict which reference points will be adopted under which conditions. The test that we provide in this paper provides a useful tool for use in that pursuit.

Related to this point, and in closing, we draw attention to two key limitations of our test.

First, our test is currently designed to apply to cases where an individual adopts a single, exogenously manipulable reference point. While these assumptions are satisfied by most reference points studied in this literature, they are not satisfied in the model of Kőszegi & Rabin (2006). However, the stochastic, endogenous reference points of that model can be accommodated with simple extensions to our framework—for example, by mixing our fixed base gambles with a third degenerate gamble to generate the equivalent of variation in the reference point, while continuing to shift all gamble payoffs by a uniform shifter Δ. We present our framework in the context of testing deterministic, exogenously manipulable reference points both because the greater simplicity of this environment makes our insights regarding identification easier to convey, but also because the vast majority of empirical estimates of reference-dependent models fit into this framework (despite the significant theoretical attention direct at the Kőszegi & Rabin (2006) model).[8]

Second, our model tests whether *all* individuals treat a specific variable as their reference point. If, for example, half of individuals adopted goal referents and half adopted average referents, then with sufficient power our test would reject either proposed reference point. This is appropriate: our test is a means of rejecting incorrect models of the reference point, and a model that does not reflect existing heterogeneity is indeed incorrect. However, the most useful version of a test would allow richly heterogeneous models to be tested. If reference-point adoption varies by observable factors, our test is directly applicable. In the example just posed, the analyst merely needs to code a new variable that takes the value of the goal referent for "goal types" and takes the value of the average referent for "average types." If, however, reference-point adoption varies by unobservable factors, our test requires

[7] Although our results are more concordant with some experiments casting doubt on expectations as reference points (see, e.g., Heffetz & List, 2014; Heffetz, 2021).

[8] Supporting this claim, the 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.

modification to be most generally applicable. In such a case, we believe that the methods we use can be productively incorporated into a mixture-modeling approach; initial examinations of such an approach appear promising.

In summary, while our overall approach does have important limitations, we believe it is relatively immediately applicable to a substantial fraction of existing reference-dependent research, and that there is promise the further development of this approach could be relevant to the full range of cases currently considered in the literature. As such, we view it as the natural first step in the development of a robust econometric theory of reference point testing and estimation, which we hope will develop hand-in-hand with experiments designed to empirically assess reference point formation through these methods.

References

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.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173–96.

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*.

Camerer, C. F. (2004). Prospect theory in the wild: Evidence from the field. In Camerer C., Loewenstein G., and Rabin M. (Ed.), *Advances in Behavioral Economics* (pp. 148 – 161). Russel Sage Foundation, Princeton University Press.

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 semiparametric functional forms. *Econometrica*, 64(4), 865–890.

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.

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.

Kahneman, D. & Tversky, A. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–232.

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.

Kőszegi, B. & Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, 99(3), 909–36.

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.

Pagel, M. (2017). Expectations-based reference-dependent life-cycle consumption. *The Review of Economic Studies*, 84(2), 885–934.

Pagel, M. (2018). A news-utility theory for inattention and delegation in portfolio choice. *Econometrica*, 86(2), 491–522.

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

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

Table 1: Baseline Gambles

	Sure Amount	*50-50 Values*	
Scenario Number	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 5 baseline gambles considered in our simulation and experiment.

Table 2: UAS Data: Summary Statistics

	(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	.

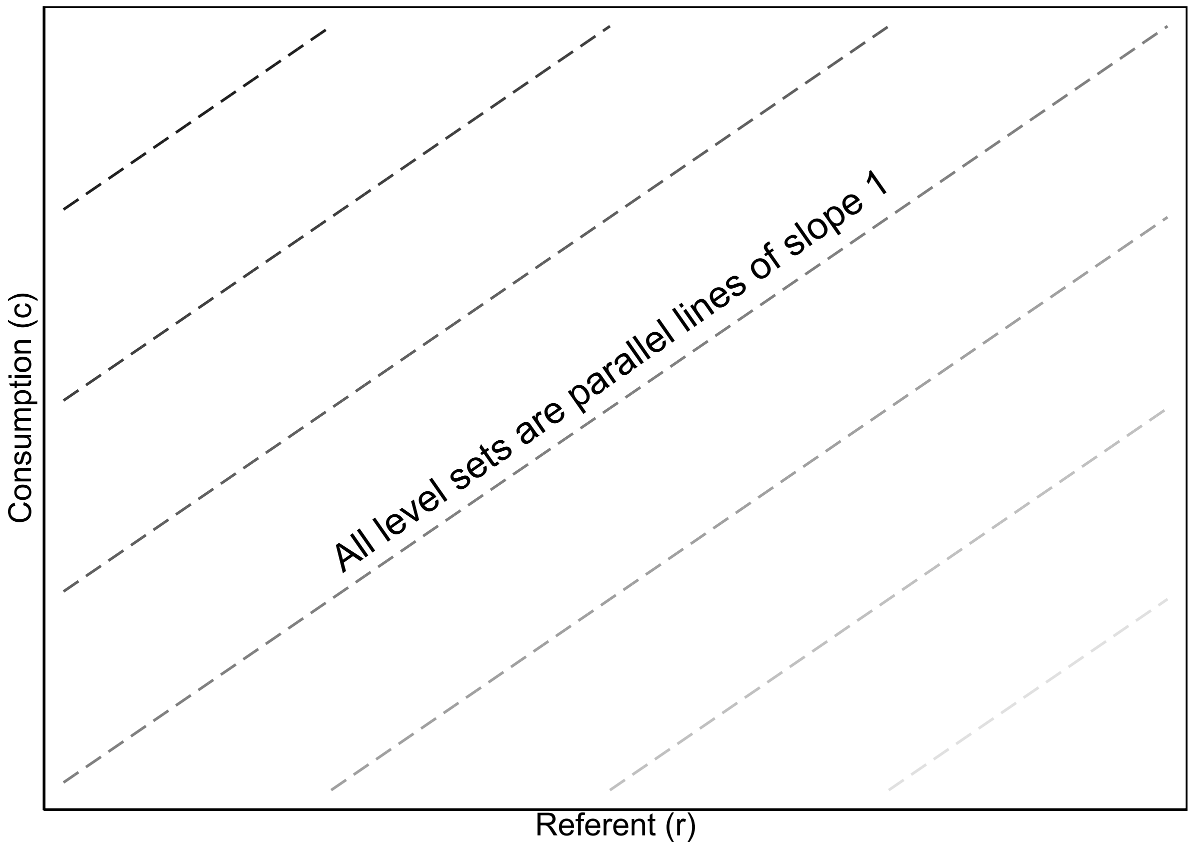
Notes: This table presents summary statistics characterizing the demographic features of our sample. With the exception of p-values, all numbers presented are the percentage of respondents with a given row's classification. The first panel characterizes a series of binary demographic variables, and the 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 3: Applying our Test in the UAS Experiment

	(1)	(2)	(3)	(4)
		p-value		
	Question Group	Stage 1	Stage 2	Passed?
Treatment: Control				
	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
	1	0.2186	0.2896	No
	2	0.232	0.904	No
Ref Pt: Average	3	0.8431	0.8728	No
	4	0.5201	0.7461	No
	5	0.1258	0.4217	No
Treatment: Goal				
	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
	1	0.1784	0.7024	No
	2	0.6404	0.8658	No
Ref Pt: Average	3	0.5201	0.9414	No
	4	0.9594	0.4757	No
	5	0.1103	0.2702	No
Treatment: Average				
	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
	1	0.0602	0.7268	No
	2	0.992	0.7059	No
Ref Pt: Average	3	0.882	0.473	No
	4	0.7511	0.5861	No
	5	0.7102	0.2785	No

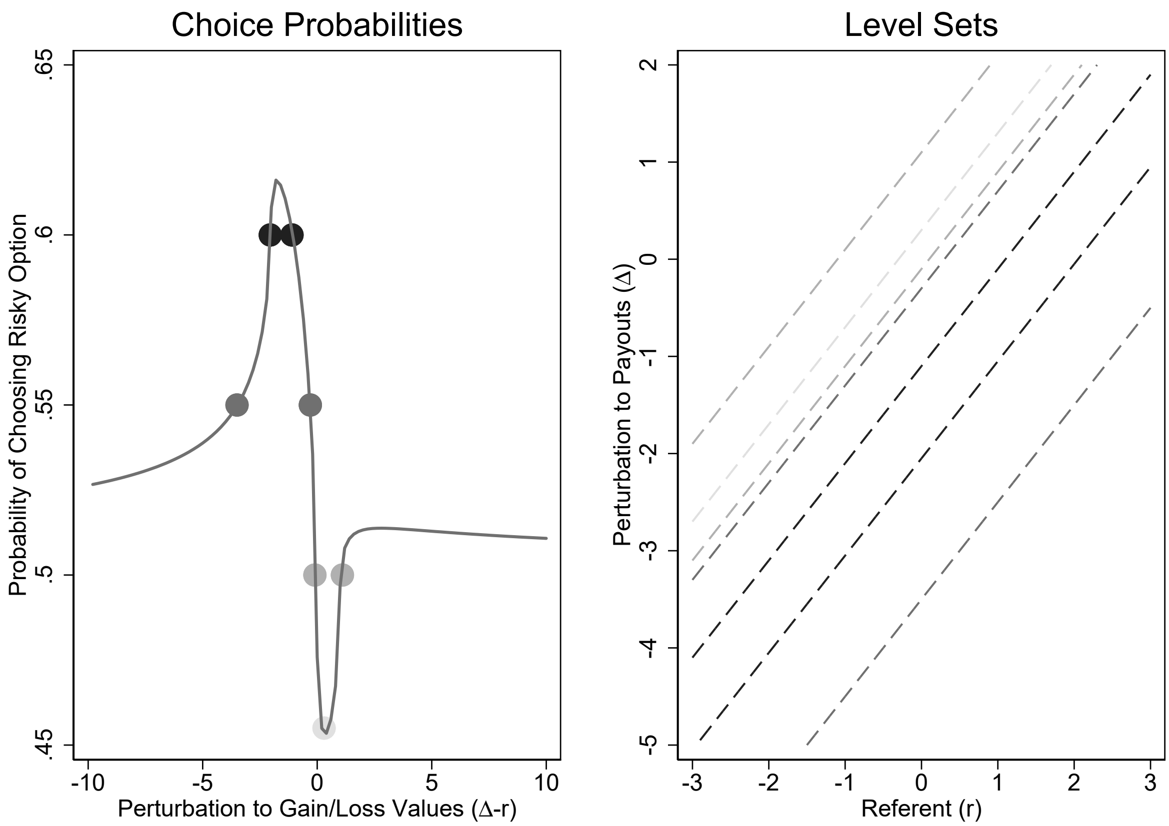
Notes: This table presents the the results from our proposed test across different treatment arms and candidate reference points. The first and second panel present results in the control arm where neither reference point is made salient. The third and fourth panel present results in the treatment arm where the goal is made salient. The fifth and sixth panel present results in the treatment arm where average earnings are made salient. Within each panel, we separately present results for each of the five question groups. The first column indexes the question 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, meaning that it has failed to reject reference dependence with respect to the candidate reference point.

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



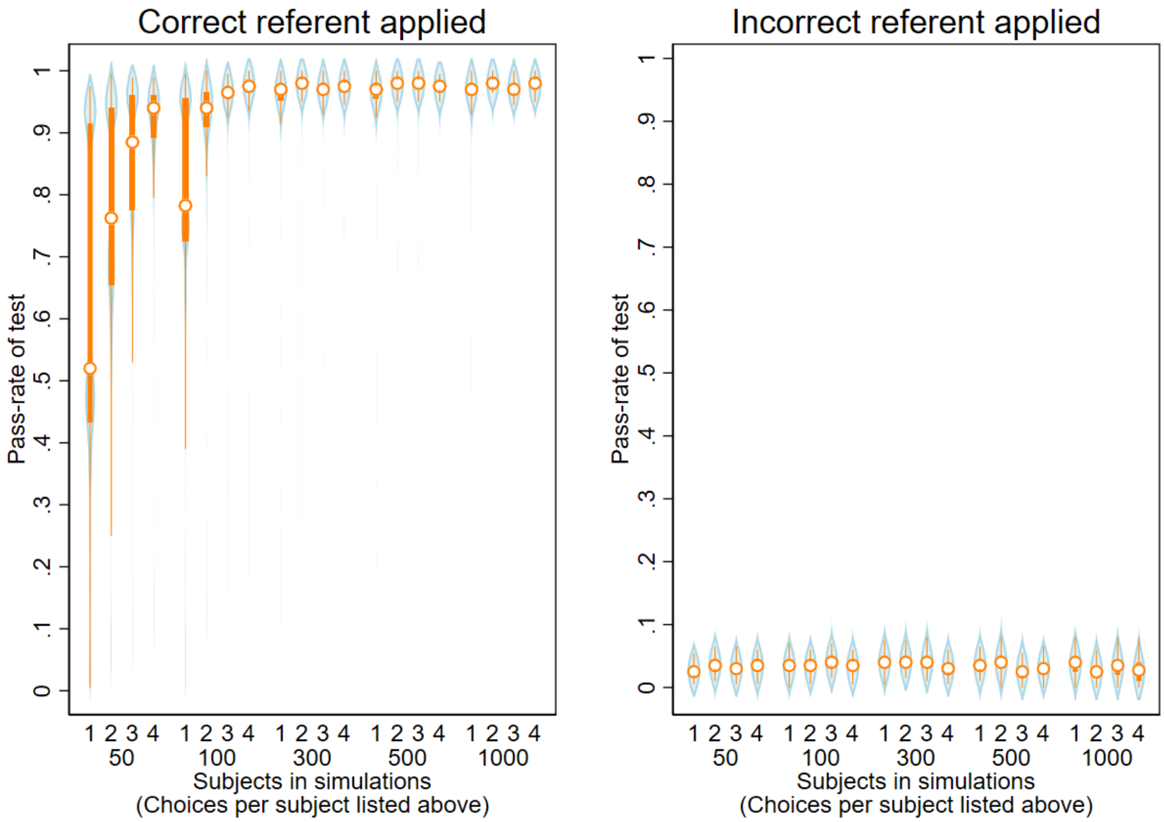
Notes: This figure represents the level sets in $C \times R$ space for a reference dependent utility function of the form $\phi(c - r)$. The dashed lines indicate example level sets plotted for this function, with darker lines denoting higher utility evaluations. At any potential consumption/referent 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.

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



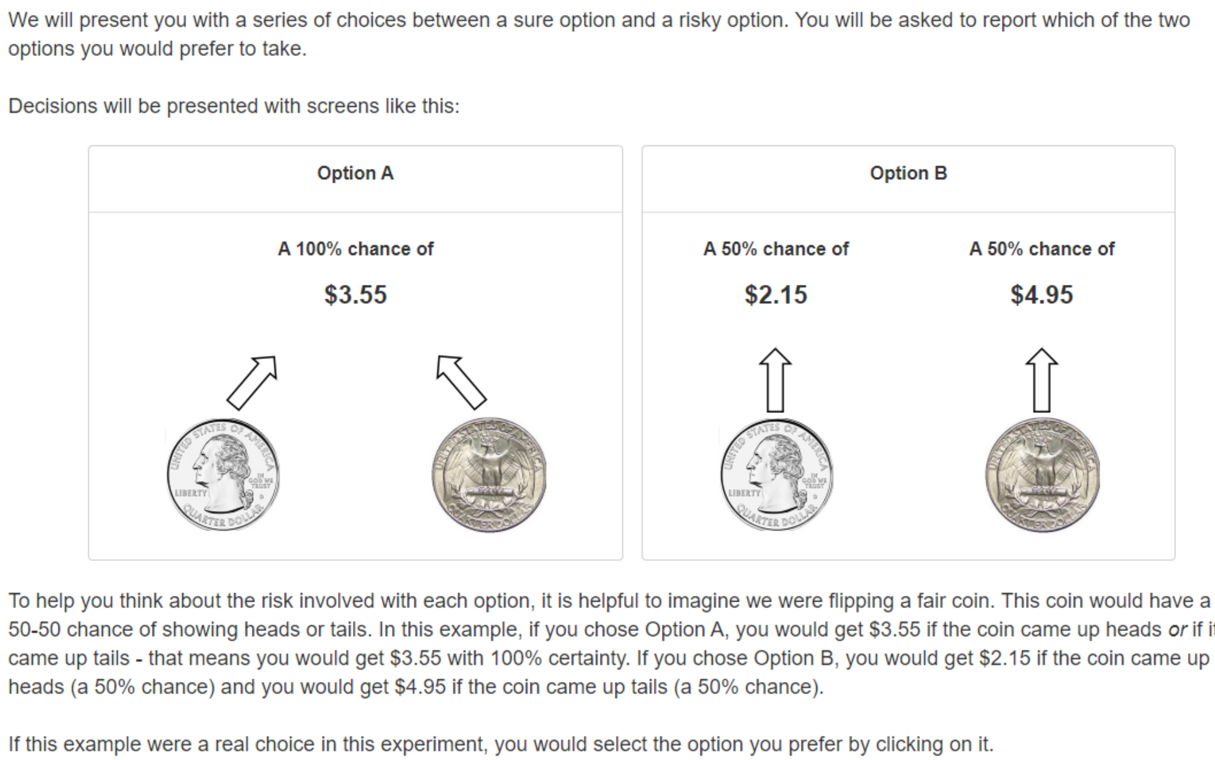
Notes: This figure presents an example choice-probability function, as well as the level-sets that arise from this function. See Section 1.2.1 for the full details of the underlying model that is simulated here.

Figure 3: Assessing Pass-Rate of Test Across Simulations



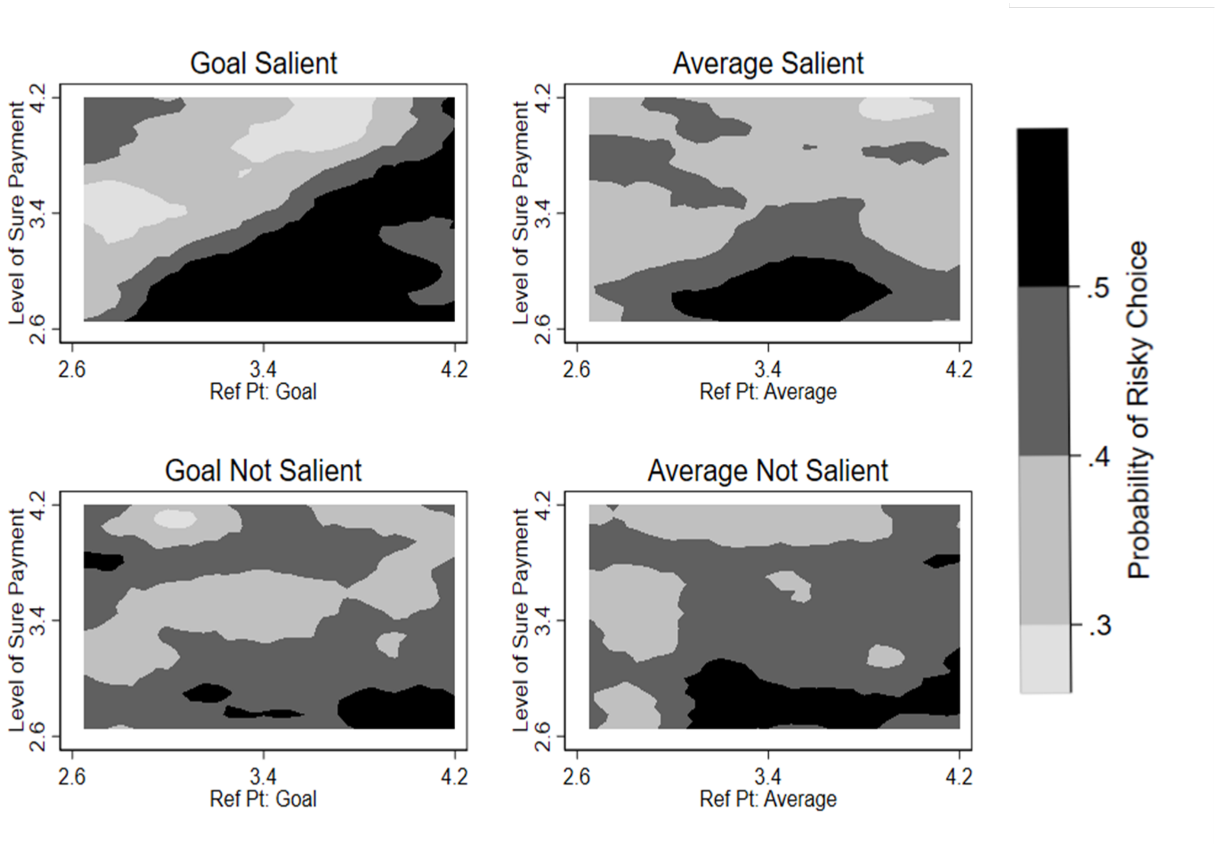
Notes: This figure summarizes our simulation study of the frequency of passing our proposed test. The left panel presents results when the test is applied to the true reference point used in the simulation—i.e., cases where the test would ideally pass. The right panel presents results when the test is applied to a 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.

Figure 4: Screenshot of Explanation of Gamble Interface



Notes: This figure presents a screenshot of the first substantive screen in our experiment. It explains the format in which gambles are presented.

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 (i.e., pooling the arm where the other reference point was salient and the arm where no reference points were salient). In this figure, we observe the parallel-line pattern that indicates a correctly specified reference point (as in Figure 2) in the top right panel, suggesting that goals are indeed used as reference points when they are made salient. In all other panels, no such pattern is observed. This suggests that we can reject that the candidate reference point was adopted in those cases, and this feature of the data drives the formal statistical rejections of these null hypotheses in Table 3. 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: $4\sigma N^{-0.35}$, where σ is the empirical standard deviation of the kernel-smoothed variable.

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 $\epsilon_{(i,j)}$. Similarly, the error term associated with predicting the binary choice with a conditional expectation function is denoted by $v_{(i,j)}$.

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_Δ and h_r to denote the bandwidth for the two dimensions respectively, and use h to denote the vector (h_Δ, h_r). 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).

Common functions Uni-dimensional kernels are denoted by $k(.)$. In case of multivariate smoothing over (Δ, r), we use the product of $k(.)$ over the individual dimensions, which is denoted by $K(.)$. Any function that are over the single dimension of $(\Delta - r)$ is indexed by a subscript or superscript α. For example, the smoothing kernel over the dimension of $(\Delta - r)$ is denoted by K^α. We will use subscripts like $K_{(i,j),(i',j')}$ to denote the pair of observations plugged in the kernel $K(.)$. The density of $\Delta_{(i,j)} - r_i$ is denoted by f_α. The joint density of (Δ, r) is denoted by $f(\Delta, r)$.

Correlation within cluster When dealing with calculations involving within-subject heterogeneity, ρ is defined such that:

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

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

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

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

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

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

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

A.2 Technical Assumptions

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

Assumption 3. *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 2. $\mathcal{K}_l$, $l \geq 1$, *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 3. $\mathcal{G}_{\mu}^{\delta}$, $\mu > 0$, $\delta > 0$, *is the class of functions* $g : R^d \to R$ *satisfying the following properties.*

(I) There exists η *such that* $\eta - 1 < \mu < \eta$ *and* g *is* $\eta - 1$ *times differentiable.*

(II) Thee exists $\rho > 0$ *such that for any* z, $sup_{y \in \{y : |y-z| < \rho\}} |g(y) - g(z) - Q_g(y, z)| / |y - z|^{\mu} \leq h_g(z)$, *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* δ*th moments.*

Definition 3 involves a slight abuse of notation: it defines the class of functions $\mathcal{G}_{\mu}^{\delta}$ despite the fact that $\mathcal{G}$ was previously used to denote gambles. We preserve this labeling to maintain consistency with Fan & Li (1996), and note that no references to gambles will occur in the course of the derivation that follows.

The next technical assumption concerns the properties of the kernels and function under null hypothesis using Definition 2 and 3. It follows Assumption A1 in Fan (1996), with minor revisions that accommodate the issue of clustering in our environment:

Assumption 4. *Any kernel in the test belongs to function class* $\mathcal{K}_2$, *and* $f_{\alpha} \in \mathcal{G}_{\lambda}^{\infty}$, *for some* $1 < \lambda \leq 2$, $g \in \mathcal{G}_{\mu}^{4}$ *where* $0 < \mu < 1$

Observation 1. *The Gaussian kernel satisfies Assumption 4.*

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 5. *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* $\eta = min(\lambda + 1, \mu)$, *where* λ *and* μ *are defined in Assumption 4.*

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 1. *If Assumption 4 and Assumption 5 hold, there exists a function* $D_g(\Delta_{(i,j)}, r_i)$ *which has fourth moment, such that as long as* $(i,j) \neq (i',j')$, $E[[g(\Delta_{(i',j')} - r_i') - g(\Delta_{(i,j)} - r_i)]k^\alpha(\frac{\Delta_{(i,j)} - \Delta_{(i',j')} - r_i + r_i'}{a})|\Delta, r] \leq D_g(\Delta_{(i,j)}, r_i)a^{1+\eta}$, *where* η *is defined in Assumption 5.*

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

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

A.3 Deriving 1st-Stage Test Statistic

Our statistic of interest is:

$$E[vf_\alpha(\Delta - r)E[vf_\alpha(\Delta - r)|\Delta, r]] \tag{17}$$

where $f_\alpha(\Delta - r)$ is the p.d.f. of $(\Delta - r)$. Thus we need to approximate v, $f_\alpha(\Delta - r)$, $f(\Delta, r)$ 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-k-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

$$\hat{g}_{(i,j)} \equiv \widehat{E}[Y_{(i,j)}|\Delta_{(i,j)} - r_i] = \frac{[(N-m)a]^{-1} \sum_{i' \neq i} \sum_{j,j'} Y_{(i',j')} K^\alpha_{(i,j),(i',j')}}{\widehat{f}_\alpha(\Delta_{(i,j)} - r_i)} \tag{18}$$

where

$$\hat{f}_{\alpha_{(i,j)}} \equiv \widehat{f}_\alpha(\Delta_{(i,j)} - r_i) = \frac{1}{(N-m)a} \sum_{i' \neq i} \sum_{j,j'} K^\alpha_{(i,j),(i',j')} \tag{19}$$

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

$$I_a = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} [\bar{v}_{(i,j)} \widehat{f}_\alpha(\Delta_{(i,j)} - r_i)][\bar{v}_{(i',j')} \widehat{f}_\alpha(\Delta_{(i',j')} - r_{i'})] K_{(i,j),(i',j')} \quad (20)$$

where

$$\bar{v}_{(i,j)} = Y_{(i,j)} - \widehat{E}[Y_{(i,j)} | \Delta_{(i,j)} - r_i] \quad (21)$$

$K_{(i,j),(i',j')}$ is a product of two univariate normal kernels where the bandwidth is h.

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

Theorem 1. *When assumption 3, 4, and 5 hold:*

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

$$\sigma_a^2 = E[f(\Delta_{(i,j)}, r_i)\sigma^4(\Delta_{(i,j)}, r_i) f_\alpha^4(\Delta_{(i,j)} - r_i)][\int k^2(s)ds]^2 \quad (22)$$

and

$$\rho_a^2 = (m^2 - 1)h(E[(\rho^4(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i)) f(\Delta_{(i,j)}, \Delta_{(i,j')}, r_i) f_\alpha^4(\Delta_{(i,j)} - r_i)] \int k^2(s)ds \quad (23)$$

(II) Under the alternative, NhI_a *converges to positive infinity with probability 1.*

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

$$T^a = \frac{NhI_a}{\sqrt{2(\hat{\sigma}_a^2 + \hat{\rho}_a^2)}} \quad (24)$$

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

$$\hat{\sigma}_a^2 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_j \sum_{j'} [\bar{v}_{(i,j)} \widehat{f}_\alpha(c_{(i,j)} - r_i)]^2 [\bar{v}_{(i',j')} \widehat{f}_\alpha(c_{(i',j')} - r_{i'})]^2 K_{(i,j),(i',j')} \int K^2(s)ds \quad (25)$$

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

$$\rho_a^2 = (m^2-1)hE[v_{(i,j_1)} f_\alpha(\Delta_{(i,j_1)} - r_i) v_{(i,j_2)} f_\alpha(\Delta_{(i,j_2)} - r_i)$$
$$E[v_{(i',j'_1)} f_\alpha(\Delta_{(i',j'_1)} - r_{i'}) v_{(i',j'_2)} f_\alpha(\Delta_{(i',j'_2)} - r_{i'}) | \Delta_{(i',j'_1)}, \Delta_{(i',j'_2)}, r_{i'}] f(\Delta_{(i',j'_1)}, \Delta_{(i',j'_2)}, r_{i'})] \int k^2(s)ds \tag{26}$$

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

$$\hat{\rho}_a^2 = \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3} \sum_i \sum_{j_1 \neq j_2} \hat{v}_{i,j_1} \hat{f}_\alpha(\Delta_{(i,j_1)} - r_i) \hat{v}_{(i,j_2)} \hat{f}_\alpha(\Delta_{(i,j_2)} - r_i) \sum_{i' \neq i} \sum_{j'_1 \neq j'_2}$$
$$\hat{v}_{(i',j'_1)} \hat{f}_\alpha(\Delta_{(i',j'_1)} - r_{i'}) \hat{v}_{(i',j'_2)} \hat{f}_\alpha(\Delta_{(i',j'_2)} - r_{i'}) K\left(\frac{\Delta_{(i,j_1)} - \Delta_{(i,j_2)}}{h}, \frac{\Delta_{(i',j'_1)} - \Delta_{(i',j'_2)}}{h}, \frac{r_{i'} - r_i}{h}\right) \int k^2(s)ds \tag{27}$$

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 I_a, 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: $f_{\alpha_{(i,j)}} = f_\alpha(\Delta_{(i,j)} - r_i)$, $g_{(i,j)} = g(\Delta_{(i,j)} - r_i)$.

The decomposition is

$$\begin{aligned} I_a = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} \{ & (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i,j)}} (g_{(i',j')} - \hat{g}_{(i',j')}) \hat{f}_{\alpha_{(i',j')}} + v_{(i,j)} v_{(i',j')} \hat{f}_{\alpha_{(i,j)}} \hat{f}_{\alpha_{(i',j')}} \\ & + \hat{v}_{(i,j)} \hat{v}_{(i',j')} \hat{f}_{\alpha_{(i,j)}} \hat{f}_{\alpha_{(i',j')}} + 2 v_{(i,j)} \hat{f}_{\alpha_{(i,j)}} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i',j')}} \\ & - 2\hat{v}_{(i,j)} \hat{f}_{\alpha_{(i,j)}} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i',j')}} - 2 v_{(i,j)} \hat{f}_{\alpha_{(i,j)}} \hat{v}_{(i',j')} \hat{f}_{\alpha_{(i',j')}} \} K_{(i,j),(i',j')} \\ \equiv {} & I_1 + I_2 + I_3 + 2I_4 - 2I_5 - 2I_6 \end{aligned} \tag{28}$$

where

$$I_1 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i,j)}} (g_{(i',j')} - \hat{g}_{(i',j')}) \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

$$I_2 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} v_{(i,j)} v_{(i',j')} \hat{f}_{\alpha_{(i,j)}} \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

$$I_3 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} \hat{v}_{(i,j)} \hat{v}_{(i',j')} \hat{f}_{\alpha_{(i,j)}} \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

$$I_4 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} v_{(i,j)} \hat{f}_{\alpha_{(i,j)}} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

$$I_5 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} \hat{v}_{(i,j)} \hat{f}_{\alpha_{(i,j)}} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

$$I_6 = \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} v_{(i,j)} \hat{f}_{\alpha_{(i,j)}} \hat{v}_{(i',j')} \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')}$$

The strategy of our proof will be to establish that, under the null, I_2 has a known asymptotic distribution and I_1, I_3, ..., I_6 all are asymptotically negligible. The following propositions establish those claims:

Proposition 3. $I_1 = o_p((Nh)^{-1})$

Proof. As discussed in Fan & Li (1996) Proposition A.1, it suffices to show that $E[I_1] = o((Nh)^{-1})$ and $E[I_1^2] = o((Nh)^{-2})$. From equation 28 we know that

$$\begin{aligned} I_1 &= \frac{1}{N(N-m)h^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i,j)}} (g_{(i',j')} - \hat{g}_{(i',j')}) \hat{f}_{\alpha_{(i',j')}} K_{(i,j),(i',j')} \\ &= \frac{1}{N(N-m)^3 h^2 a^2} \sum_i \sum_{i' \neq i} \sum_{j,j'} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{\tilde{j},\tilde{j}'} (g_{(i,j)} - g_{(\tilde{i},\tilde{j})}) K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})} (g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')}) K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')} K_{(i,j),(i',j')} \end{aligned} \tag{29}$$

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 $E[I_1] = o((Nh)^{-1})$ 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 $E[I_1^2] = o((Nh)^{-2})$ We have

$$E[I_1^2] = \frac{1}{N^2(N-m)^6h^4a^4} \sum_{i} \sum_{i' \neq i} \sum_{j,j'} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{\tilde{j},\tilde{j}'} \sum_{k} \sum_{k' \neq k} \sum_{l,l'} \sum_{\tilde{k} \neq k} \sum_{\tilde{k}' \neq k'} \sum_{\tilde{l},\tilde{l}'}$$
$$(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}(g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')})K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')}$$
$$(g_{(k,l)} - g_{(\tilde{k},\tilde{l})})K^{\alpha}_{(k,l),(\tilde{k},\tilde{l})}(g_{(k',l')} - g_{(\tilde{k}',\tilde{l}')})K^{\alpha}_{(k',l'),(\tilde{k}',\tilde{l}')}K_{(k,l),(k',l')}$$

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 $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}$ so that Lemma 1 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 1 can be applied to all terms like $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}$, 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 1 can be applied to all terms like $(g_{(i,j)} - g_{(\tilde{i},\tilde{j})})K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}$ except for $(g_{(k,l)} - g_{(\tilde{k},\tilde{l})})K^{\alpha}_{(k,l),(\tilde{k},\tilde{l})}$, 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 1 can be applied to $(g_{(i',j')} - g_{(\tilde{i}',\tilde{j}')})K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}$ and $(g_{(k,l)} - g_{(\tilde{k},\tilde{l})})K^{\alpha}_{(k,l),(\tilde{k},\tilde{l})}(g_{(k',l')} - g_{(\tilde{k}',\tilde{l}')})K^{\alpha}_{(k',l'),(\tilde{k}',\tilde{l}')}$, 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 4. *$NhI_2 \to N(0, 2(\sigma_a^2 + \rho_a^2))$ in distribution, where*

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

and

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

where k(.) is the Gaussian kernel.

Proof. Consider term I_2. We have

$$\begin{aligned} I_2 &= \frac{1}{N(N-m)^3h^2a^2} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')} \\ &= \frac{1}{N(N-m)^3h^2a^2} \sum_{i \neq i' \neq \tilde{i} \neq \tilde{i}'} \sum_{j,j',\tilde{j},\tilde{j}'} v_{(i,j)}v_{(i',j')}K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')} \\ &\quad + I_2R \\ &= I_2U + I_2R \end{aligned} \tag{30}$$

Here the terms where ii, i', $\tilde{i}$, $\tilde{i}'$ do not equal each other is denoted by I_2U. 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-k-out estimator, where the cross products of some dependent observations are omitted, addresses this issue.

To see this formally, rewrite I_2U in terms of U-statistics:

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

where

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

and

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

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

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

$$\begin{aligned}E[P_n(\mathcal{Z}_i,\mathcal{Z}_{i'})^2]&=16E[(\sum_{j,j'}v_{(i,j)}v_{(i',j')}K_{(i,j),(i',j')}\sum_{\tilde{j},\tilde{j}'}E[K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]\\&=16m^4E[(\sum_{j,j'}v_{(i,j)}v_{(i',j')}K_{(i,j),(i',j')}E[K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]\\&=16m^6E[v^2_{(i,j)}v^2_{(i',j')}K^2_{(i,j),(i',j')}(E[K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]+\\16m^6(m^2-1)E[v_{(i,j)}v_{(i',j')}v_{(i,j_*)}v_{(i',j'_*)}&K_{(i,j),(i',j')}K_{(i,j_*),(i',j'_*)}(E[K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}])^2]\end{aligned} \tag{31}$$

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

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

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

$$
\begin{aligned}
&16m^6(m^2-1)E[v_{(i,j)}v_{(i',j')}v_{(i,j_*)}v_{(i',j'_*)}K_{(i,j),(i',j')}K_{(i,j_*),(i',j'_*)}(E[K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i}',\tilde{j}')}|\mathcal{Z}_i,\mathcal{Z}_{i'}]))^2] \\
&= 16m^6(m^2-1)a^4E[\rho^2(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)\rho^2(\Delta_{(i',j')},\Delta_{(i',j'_*)},r_{i'})K_{(i,j),(i,j_*)}K_{(i',j'),(i',j'_*)} \\
&\quad (\int K^w(u)K^w(v)f_\alpha(\Delta_{(i,j)}-r_i+au)f_\alpha(\Delta_{(i',j')}-r_{i'}+av)dudv)^2] \\
&= 16m^6(m^2-1)a^4h^3\int \rho^2(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)\rho^2(\Delta_{(i,j)}+hs_1,\Delta_{(i,j_*)}+hs_2,r_i+hs_3)f(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i) \\
&\quad f(\Delta_{(i,j)}+hs_1,\Delta_{(i,j_*)}+hs_2,r_i+hs_3)k(s_1)k(s_2)k^2(s_3)(\int k^\alpha(u)k^\alpha(v)f_\alpha(\Delta_{(i,j)}-r_i+av) \\
&\quad f_\alpha(\Delta_{(i,j)}+hs_1-r_i-hs_2+au)dudv)^2ds_1ds_2ds_3d\Delta_{(i,j)}d\Delta_{(i,j_*)}dr_i \\
&= 16m^6(m^2-1)a^4h^3(E[(\rho^4(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i))f(\Delta_{(i,j)},\Delta_{(i,j_*)},r_i)f_\alpha(\Delta_{(i,j)}-r_i)^4]\int k^2(s_3)ds_3+o(1))
\end{aligned}
$$

where $s_1\equiv\frac{\Delta_{i',j'}-\Delta_{i,j}}{h}$, $s_2\equiv\frac{\Delta_{i,j'_*}-\Delta_{i,j_*}}{h}$, $s_3\equiv\frac{r_{i'}-r_i}{h}$, $u\equiv\frac{(\Delta_{(\tilde{i},\tilde{j})}-r_{\tilde{i}})-(\Delta_{i,j}-r_i)}{a}$, $v\equiv\frac{(\Delta_{(\tilde{i}',\tilde{j}')}-r_{\tilde{i}'})-(\Delta_{i',j'}-r_{i'})}{a}$

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

$$NhI_2U\xrightarrow{d}N(0,2(\sigma_a^2+\rho_a^2))$$

where

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

Finally we have,

$$
\begin{aligned}
I_2R=\frac{1}{N(N-m)^3h^2a^2}\{&\sum_{i\neq i'\neq\tilde{i}}\sum_{j,j',\tilde{j},\tilde{j}'}v_{(i,j)}v_{(i',j')}K^w_{(i,j),(\tilde{i},\tilde{j})}K^w_{(i',j'),(\tilde{i},\tilde{j}')}K_{(i,j),(i',j')}+ \\
&\sum_{i\neq i'\neq\tilde{i}}\sum_{j,j',\tilde{j},\tilde{j}'}v_{(i,j)}v_{(i',j')}K^w_{(i,j),(i',\tilde{j})}K^w_{(i',j'),(\tilde{i},\tilde{j}')}K_{(i,j),(i',j')}+ \\
&\sum_{i\neq j}\sum_{j,j',\tilde{j},\tilde{j}'}v_{(i,j)}v_{(i',j')}K^w_{(i,j),(i',\tilde{j})}K^w_{(i',j'),(i,\tilde{j}')}K_{(i,j),(i',j')}\} \\
&=I_2R^1+I_2R^2+I_2R^3
\end{aligned}
$$

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

$$\begin{aligned} NhI_2R^1 &= \frac{\binom{n}{3}}{(N-m)^3ha^2}[\binom{n}{3}^{-1} I_2R^1] \\ &= n^{-1}\frac{\binom{n}{3}}{(N-m)^3ha^2}[n\binom{n}{3}^{-1} I_2R^1] \end{aligned}$$

As m is a constant, and $n\binom{n}{3}^{-1} I_2R^1 = o_p(ah^2)$, we have

$$NhI_2R^1 = o_p(\frac{1}{na}) = o_p(1)$$

Similarly we have

$$NhI_2R^2 = o_p(\frac{1}{na}) = o_p(1)$$

and

$$NhI_2R^3 = o_p(\frac{1}{na}) = o_p(1)$$

So now we can conclude that

$$NhI_2 \to N(0, 2(\sigma_a^2 + \rho_a^2))$$

□

Proposition 5. $I_3 = o_p((Nh)^{-1})$

Proof. From equation 28, we have,

$$I_3 = \frac{1}{N(N-m)^3h^2a^2}\sum_{i}\sum_{i'\neq i}\sum_{\tilde{i}\neq i}\sum_{\tilde{i}'\neq i'}\sum_{j,j',\tilde{j},\tilde{j}'} v_{(\tilde{i},\tilde{j})}v_{(\tilde{i}',\tilde{j}')}K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')}$$
$$\equiv I_3F + I_3S$$

where I_2F is the sum of the terms where $\tilde{i} = \tilde{i}'$. I_2S is the sum of the terms where $\tilde{i} \neq \tilde{i}'$.

For I_2F, we have

$$E[I_3F] = \frac{1}{N(N-m)^3h^2a^2} * O(\sum_{i}\sum_{i'\neq i}\sum_{\tilde{i}\neq i}\sum_{\tilde{i}'\neq i'}\sum_{j,j',\tilde{j},\tilde{j}'} E[v_{(\tilde{i},\tilde{j})}v_{(\tilde{i},\tilde{j}')}K^{\alpha}_{(i,j),(\tilde{i},\tilde{j})}K^{\alpha}_{(i',j'),(\tilde{i}',\tilde{j}')}K_{(i,j),(i',j')}])$$
$$= O(\frac{1}{na}) = o(\frac{1}{nh})$$

and

$$E[I_3F^2] = o(\frac{1}{(na)^2}) = o(\frac{1}{(nh)^2})$$

$$I_3S = I_3SF + I3_SS$$

where I_3SF is the sum of the terms where i, i', $\tilde{i}$, $\tilde{i}'$ take different values. I_3SS is the sum of the terms where at least two of the values equal each other.

Regarding I_3SF, we have,

$$E[I_3SF^2] = \frac{1}{n^8h^4a^4} * o(n^6h^4 * a * 2) * max\{o(\frac{1}{nh}, \frac{1}{na}, 1)\}$$
$$= o(\frac{1}{n^2a^2}) = o(\frac{1}{n^2h^2})$$

Regarding I_3SS, we have,

$$E[|I_3SS|] = \frac{1}{N^4h^2a^2} * O(N^3ah^2) = O(\frac{1}{na}) = o(\frac{1}{nh})$$

□

Proposition 6. $I_4 = o_p((Nh)^{-1})$

Proof. We have

$$I_4 = \sum_i \sum_{i' \neq i} v_{(i,j)} \sum j, j' \hat{f}_{\alpha_{(i,j)}} (g_{(i,j)} - \hat{g}_{(i,j)}) \hat{f}_{\alpha_{(i',j')}}$$

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

Since

$$E[I_4] = 0$$

and

$$E[I_4^2] = \frac{1}{N^2(N-m)^6 h^4 a^4} \sum_i \sum_{i' \neq i} \sum_{\tilde{i} \neq i} \sum_{\tilde{i}' \neq i'} \sum_k \sum_{k' \neq k} \sum_{\tilde{k} \neq k} \sum_{\tilde{k}' \neq k'} \sum_{j,j',\tilde{j},\tilde{j}',l,l',\tilde{l},\tilde{l}'}$$

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

$$= I_4 P + I_4 R$$

where $I_4 P$ is the sum of the terms where $i = k$ and $i, i', \tilde{i}, \tilde{i}', k', \tilde{k}, \tilde{k}'$ are pairwise different. $I_4 R$ represents all the other terms. $I_4 P$ is of order $O(n^{-1} a^{2\eta}) = o(\frac{1}{n^2 h}) = o(\frac{1}{n^2 h^2})$. When two of these indices equal each other, $I_4 R$ is at most of order $O(n^{-1} a^{2\eta} max\{\frac{a}{n a^{\eta}}, \frac{1}{n a^{\eta}}, \frac{1}{n h^2}\}) = o(\frac{1}{n^2 h^2})$ □

Proposition 7. $I_5 = o_p((Nh)^{-1})$

Proof. As I_5 is algebraically of the same structure as I_4, the derivation is the same as Proposition 6. □

Proposition 8. $I_6 = o_p((Nh)^{-1})$

Proof. We have:

$$I_6 = \sum_i \sum_{i' \neq i} \sum_{j,j'} v_{(i,j)} \hat{f}_{\alpha_{(i,j)}} \hat{v}_{(i',j')} \hat{f}_{\alpha_{(i',j')}} \} K_{(i,j),(i',j')}$$

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

$$\equiv I_6 F + I_6 S$$

where I_6F denotes the sum the of terms where $\tilde{i}' = i$, I_6S denotes the sum of the terms where $\tilde{i}' \neq i$. We have, Similar to Proposition A.6 in Fan & Li (1996), we have $E[(I_6F)^2] = o(n^{-2}h^{-2})$ and $E[(I_6S)^2] = o(n^{-2}h^{-2})$. Thus $I_6 = o_p((Nh)^{-1})$. □

Proposition 9. $\hat{\sigma}_a^2 = \sigma_a^2 + o_p(1)$ *and* $\hat{\rho}_a^2 = \rho_a^2 + o_p(1)$

Proof. We know from equation 25 that

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

using Lemma 1 and discussions in Proposition 3 to Proposition 8, we have,

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

Similarly,

$$\hat{\rho}_a^2 = \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}\hat{v}_{i,j_1}\hat{f}_\alpha(\Delta_{(i,j_1)}-r_i)\hat{v}_{(i,j_2)}\hat{f}_\alpha(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$\hat{v}_{(i',j_1')}\hat{f}_\alpha(\Delta_{(i',j_1')}-r_{i'})\hat{v}_{(i',j_2')}\hat{f}_\alpha(\Delta_{(i',j_2')}-r_{i'})K(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h},\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h},\frac{r_{i'}-r_i}{h})\int k^2(s)ds$$
$$= \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}v_{i,j_1}\hat{f}_\alpha(\Delta_{(i,j_1)}-r_i)v_{(i,j_2)}\hat{f}_\alpha(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$v_{(i',j_1')}\hat{f}_\alpha(\Delta_{(i',j_1')}-r_{i'})v_{(i',j_2')}\hat{f}_\alpha(\Delta_{(i',j_2')}-r_{i'})K(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h},\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h},\frac{r_{i'}-r_i}{h})\int k^2(s)ds+o_p(1)$$
$$= \frac{(m^2-1)h}{N(N-m)(m-1)^2h^3}\sum_i\sum_{j_1\neq j_2}v_{i,j_1}f_\alpha(\Delta_{(i,j_1)}-r_i)v_{(i,j_2)}f_\alpha(\Delta_{(i,j_2)}-r_i)\sum_{i'\neq i}\sum_{j_1'\neq j_2'}$$
$$v_{(i',j_1')}f_\alpha(\Delta_{(i',j_1')}-r_{i'})v_{(i',j_2')}f_\alpha(\Delta_{(i',j_2')}-r_{i'})K(\frac{\Delta_{(i,j_1)}-\Delta_{(i,j_2)}}{h},\frac{\Delta_{(i',j_1')}-\Delta_{(i',j_2')}}{h},\frac{r_{i'}-r_i}{h})\int k^2(s)ds+o_p(1)$$
$$= \hat{\rho}_a^2 = \rho_a^2 + o_p(1) \tag{32}$$

□

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 3, 4, and 5 hold:*

(I) Under the null, $NhI_a \rightarrow N(0, 2(\sigma_a^2+\rho_a^2))$, where

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

and

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

(II) Under the alternative, NhI_a converges to positive infinity with probability 1.

Proof. To establish part (I), recall from equation 20 that we can decompose the test statistics into

$$I_1 + I_2 + I_3 + 2I_4 - 2I_5 - 2I_6$$

Proposition 4 establishes that I_2 has exactly the distribution noted in the theorem. Propositions 3, 5, 6, 7, and 8 establish that all other terms are asymptotically negligible. Proposition 9 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 o f 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,

$$\hat{\mu} = \frac{1}{N} \sum_i \sum_{m_i} Y_{(i,m_i)} \tag{33}$$

We base our test on the the following condition:

$$E[uf(x)E[u|x]] = 0 \tag{34}$$

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

$$
\begin{aligned}
I_\mu &= \frac{1}{N(N-m)a}\sum_i\sum_{i'\neq i}\sum_{j,j'}(Y_{(i,j)}-\hat{\mu})(Y_{(i',j')}-\hat{\mu})K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) \\
&= \frac{1}{N(N-m)a}\sum_i\sum_{i'\neq i}\sum_{j,j'}(u_{(i,j)}+\mu-\hat{\mu})(u_{(i',j')}+\mu-\hat{\mu})K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) \\
&= (\mu-\hat{\mu})^2\frac{1}{N(N-m)a}\sum_i\sum_{i'\neq i}\sum_{j,j'}K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) \\
&+(\mu-\hat{\mu})\frac{1}{N}\sum_i\sum_j u_{(i,j)}\frac{1}{(N-m)a}\sum_{i'\neq i}\sum_{j'}K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) \\
&+(\mu-\hat{\mu})\frac{1}{N}\sum_{i'}\sum_{j'} u_{(i',j')}\frac{1}{(N-m)a}\sum_{i\neq i'}\sum_{j}K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) \\
&+\frac{1}{N(N-m)a}\sum_i\sum_{i'\neq i}\sum_{j,j'}u_{(i,j)}u_{(i',j')}K(\frac{X_{(i',j')}-X_{(i,j)}}{a})
\end{aligned}
\tag{35}
$$

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.

$$
\begin{aligned}
N\sqrt{a}I_\mu &= \frac{N\sqrt{a}}{N(N-m)a}\sum_i\sum_{i'\neq i}\sum_j\sum_{j'}u_{(i,j)}u_{(i',j')}K(\frac{X_{(i',j')}-X_{(i,j)}}{a}) + o_p(1) \\
&= \frac{\binom{n}{2}m}{N(N-m)\sqrt{a}}\{n\binom{n}{2}^{-1}\sum_{1\leq i<j\leq n}[2\sum_j\sum_{j'}u_{(i,j)}u_{(i',j')}K(\frac{X_{(i',j')}-X_{(i,j)}}{a})]\} + o_p(1) \\
&\rightarrow N(0, 2(\sigma_\mu^2+\rho_\mu^2))
\end{aligned}
\tag{36}
$$

where

$$\sigma_\mu^2 = E[u^2]^2E[f(X)]\int K^2(u)du \tag{37}$$

$$\rho_\mu^2 = E[\rho^4(X_{(i,j)}, X_{(i',j')})f(X_{(i,j)}, X_{(i',j')})] \tag{38}$$

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

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

$$\sigma_a^2 = E[u^2]^2 E[f(X)] \int K^2(u)du$$

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

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

The estimator for σ_a^2 and ρ_a^2 is

$$\hat{\sigma}_\mu^2 = \frac{1}{N(N-m)a} \sum_i \sum_j \hat{u}_{(i,j)}^2 \sum_{i' \neq i} \sum_{j'} \hat{u}_{(i',j')}^2 k(\frac{X_{(i',j')} - X_{(i,j)}}{a}) \int k^2(u)du \tag{39}$$

$$\begin{aligned} \hat{\rho}_\mu^2 = \frac{(m^2-1)a}{N(N-m)(m-1)^2a^2} & \sum_i \sum_{j_* \neq j} \hat{u}_{(i,j)}\hat{u}_{(i,j_*)} \sum_{i' \neq i} \sum_{j'_* \neq j'} \hat{u}_{(i',j')}\hat{u}_{(i',j'_*)} \\ & k(\frac{X_{(i',j')} - X_{(i,j)}}{a})k(\frac{X_{(i',j'_*)} - X_{(i,j_*)}}{a}) \end{aligned} \tag{40}$$

and the test statistics is

$$\frac{N\sqrt{a}I_\mu}{\sqrt{2(\hat{\sigma}_\mu^2 + \hat{\rho}_\mu^2)}} \tag{41}$$

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