Visual Inference and Graphical Representation in Regression Discontinuity Designs

Christina Korting
University of Delaware

Zhuan Pei*
Cornell University and IZA

Carl Lieberman
Princeton University

Yi Shen
University of Waterloo

Jordan Matsudaira
Columbia University

October 14, 2020

Abstract

Despite the widespread use of graphs in empirical research, little is known about readers' ability to process the statistical information they are meant to convey ("visual inference"). We evaluate visual inference in regression discontinuity (RD) designs by measuring how accurately readers identify discontinuities in graphs produced from data generating processes calibrated on 11 published papers. First, we assess the effects of different graphical representation methods on visual inference using randomized experiments, finding that bin widths and fit lines have the largest impacts. Second, we evaluate visual inference by experts and non-experts, finding that experts perform comparably to non-experts but only partly predict the effects of different graphical methods. Third, we use experts' visual inference as a benchmark to evaluate common econometric procedures in RD designs, finding that visual inference achieves similar or lower type I error rates. Fourth, we conduct an eyetracking study on RD visual inference, but find no gaze patterns that robustly predict successful inference. We also study visual inference in the related regression kink design.

Key Words: Regression Discontinuity Design; Regression Kink Design; Graphical Methods; Visual Inference; Eyetracking; Expert Prediction

JEL Code: A11, C10, C40

*Corresponding author. Assistant Professor, Department of Policy Analysis and Management, Cornell University, Ithaca, NY 14853, USA; zhuan.pei@cornell.edu

We have benefited from comments by Colin Camerer, Matias Cattaneo, Damon Clark, Nathan Grawe, David Lee, Lars Lefgren, Thomas Lemieux, Pauline Leung, Jia Li, Adam Loy, Alex Mas, Doug Miller, Ted O'Donoghue, Steve Pischke, Jesse Rothstein, Rocio Titiunik, Stephanie Wang, and Andrea Weber, as well as seminar participants at Cornell, UC Irvine, and Princeton. Michael Daly, Rebecca Jackson, Motasem Kalaji, Xingyue Li, Fiona Qiu, and Tatiana Velasco provided research assistance, and we thank Brad Turner and Patti Tracey for providing logistical support. We are indebted to our friends for testing the experiment, to colleagues for participating in the study, and to Sahara Byrne for giving us access to her eyetracking lab. Financial support from the Cornell Institute of Social Sciences and the Princeton Industrial Relations Section is gratefully acknowledged. The authors have no relevant or material financial interests that relate to the research of this paper. This study is registered in the Open Science Framework and the AEA RCT Registry with ID AEARCTR-0004331.

1 Introduction

"Few would deny that the most powerful statistical tool is graph paper."

— Geoffrey S. Watson (1964)

Despite the widespread use of graphical analysis in applied economic studies, there is little research on readers' ability to process the statistical information graphs are meant to convey (visual statistical inference or *visual inference* per Majumder, Hofmann, and Cook, 2013), especially in the empirical applications where graphs are most commonly used. In this paper, we evaluate the performance of visual inference in the regression discontinuity design (RDD or RD design). The popularity of RDD in the modern causal inference toolkit, which began in economics with Angrist and Lavy (1999), makes it an important setting in which to study visual inference. Standard practices in applying RDD today perhaps best embody the spirit of Watson's quote above, with graphs playing a central role in the presentation of findings. The key RD graph plots the bivariate relationship between outcome variable Y and running variable X and is meant to display a discontinuity (or lack thereof) in the underlying conditional expectation function (CEF) as X crosses a policy threshold. Influential practitioner guides by Imbens and Lemieux (2008) and Lee and Lemieux (2010) recommend creating this graph by dividing X into bins, computing the average of Y within each bin, and generating a scatter plot of these Y-averages against the midpoints of the bins.

We assess the performance of visual inference by studying whether people presented with this graph can correctly identify the existence or absence of a discontinuity at the policy threshold. This endeavor raises four conceptual and practical questions. First, because we can construct many different RD graphs based on the same data, how do graphical representation techniques affect visual inference? Second, as readers may have different abilities, whose visual inference should we evaluate? Third, what benchmark should we use to judge the adequacy of visual inference? And fourth, with the hope of improving it, can we understand the mechanics of visual inference?

To address the first question, we build on work pioneered by Eells (1926) and refined by Cleveland and McGill (1984a) who experimentally evaluate the accuracy of visual perception across a variety of perceptual tasks (such as discerning the area, angle, or saturation of a graphical element). We conduct a series of randomized experiments to examine how different graphical parameters affect visual inference in RD. We present participants with RD graphs produced from data generating processes (DGPs) based on microdata from 11 published papers that we randomly selected from a list of 110 empirical studies drawn from top

economics journals (the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal of Business and Economic Statistics*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics and Statistics*). For each graph, we ask participants to identify the existence or absence of a discontinuity. We randomize respondents into different treatment arms and show participants within each arm graphs produced with particular graphical parameters such as small bin widths and evenly spaced bins.

There is limited research on how to choose these parameters in practice. For example, Calonico, Cattaneo, and Titiunik (2015) propose two bin width selectors: one that minimizes the integrated mean squared error (IMSE) of the bin averages, resulting in fewer, larger bins, and another that mimics the variability of the underlying data (mimicking variance or MV), which leads to more, smaller bins. While both proposals set a graphical parameter to satisfy an econometric criterion, practitioners are left with little basis to choose between them. Moreover, a host of other choices over graphical parameters remains with minimal guidance from the literature, such as including smoothed regression lines in the binned scatter plot, adding a vertical line to indicate the policy threshold, and choosing the axis scales.

By comparing the rates at which respondents correctly classify discontinuities across treatment arms, we can assess the pros and cons of different graphical parameters, which informs best practices in RD visual representation. We find that certain graphical parameters such as bin width and imposing smoothed regression lines create important tradeoffs between type I error (incorrectly identifying a discontinuity when there is none) and type II error (incorrectly identifying the absence of a discontinuity when there is one) rates. Relative to IMSE bins and imposing fit lines, using MV bins and not including the fit lines tends to decrease type I error rates but increase type II error rates.

To address the second question of evaluating visual inference and to assess how our estimates generalize across experience levels, we compare the performance of non-experts in our randomized experiments recruited through the Cornell University Johnson College's Business Simulation Lab with the performance of experts in fields that often employ RD designs. We find that the two groups perform comparably and are similarly affected by graphical methods. In the spirit of DellaVigna and Pope (2018), we also test whether experts are able to predict the graphical parameters that result in the highest rate of visual inference success by non-experts. We find that experts only partly anticipate the aforementioned tradeoffs from bin widths and fit lines.

Third, as a benchmark for the performance of visual inference, we compare the type I and type II error

rates of our crowdsourced discontinuity/kink classifications with those from inference based on econometric estimators at the 5% (asymptotic) level. These comparisons confirm that the graphical representation of RD designs is a valuable tool for empirical analysis and complements econometric inference. We find that visual inference on graphs generated with methods leading to the lowest type I error rates performs very similarly to the Armstrong and Kolesár (2018) RD inference procedure (henceforth AK). RD visual inference achieves lower type I error rates than tests based on the Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) estimators (henceforth IK and CCT, respectively), but the two procedures exhibit considerable power advantages. The examination of the joint distribution of visual and econometric tests further illustrates their complementarity: while they commit similar type II errors, there does not appear to be a strong association in their type I errors.

Fourth, to understand the mechanics behind successful visual inference and investigate avenues for its improvement, we conduct an eyetracking study with both expert and non-expert participants. We track participants' visual attention to different parts of RD graphs while they form their discontinuity assessments. Although the six experts outperform the nine non-experts in this small lab study, their eyegaze patterns are similar along most dimensions, and various prediction exercises do not reveal eyegaze measures that robustly correlate with visual inference success. We tentatively conclude that the *processing* of visual signals, as opposed to their acquisition through ocular movements, is likely the driver of performance heterogeneity in visual inference.

We also evaluate visual inference in the closely related regression kink design (RK design or RKD). First, we use an analogous experimental approach to study graphical representation and additionally test the effects of visualization methods specific to RKD, namely plotting detrended outcomes as in Lee et al. (2019) and plotting first derivative estimates by adapting the generalized bin selector from Cattaneo, Farrell, and Feng (2020). Like in RD, we find similar tradeoffs in the choice of bin width, while plotting detrended outcomes decreases type II error rates and plotting derivative estimates leads to poor visual inference performance. Second, we document the performances of both experts and non-experts and observe that experts attain similar type I error rates but lower type II error rates compared with non-experts. Third, both experts and non-experts struggle to differentiate kinks from smooth peaks, which the default CCT inference procedure is better able to handle. However, the latter has a serious power disadvantage even at the largest kink magnitude we test. The lack of power can be alleviated by using conventional inference methods based on a local linear or quadratic estimator in conjunction with a larger bandwidth, consistent with the findings by Card, Lee,

Pei, and Weber (2017).

The results from this paper lead to takeaways of RD/RK graphical representation and econometric inference for both practitioners and econometricians. For graphical representation in RD, if the primary concern is type I error rates, consistent with the notion of size control in econometric inference, one should use small bins with no fit lines as opposed to large bins with fit lines. Other parameters such as bin spacing (equally spaced versus quantile-spaced), axis scaling, and the presence of a vertical line indicating the policy threshold do not appear to matter, implying that researchers can adhere to reasonable personal preferences. For RK, one may consider plotting detrended outcomes, while plotting derivative estimates is not recommended. The RD bin size result particularly underscores the value of using human perception as the criterion for evaluating graphs: while the larger IMSE bins are derived from a conventional statistical criterion, they have important drawbacks compared to the MV bins, which come from a non-standard statistical criterion and are constructed in a more ad-hoc way.

For econometric inference, the performance of visual inference provides an important benchmark: it informs the econometric procedure's value added above what readers glean from a carefully constructed graph. We find that the experts' performance is almost identical to that of the AK RD inference. By comparison, the IK and CCT RD procedures have elevated type I error rates (the CCT type I error rates are not significantly higher) but lower type II error rates. We can lower their type I error rates to that of the experts (just below 8%) by increasing the critical t-values to 2.46 for IK and 2.28 for CCT, and the resulting inference procedures retain their power advantages. Econometric methods add the most value in delivering estimates of the discontinuity magnitude: for example, the IK estimator yields lower mean squared errors than experts across all 11 DGPs.

This paper brings together a diverse set of literatures and makes the following contributions. i) We are the first to evaluate visual inference and graphical representation practices in two widely used quasi-experimental research designs in economics and other social sciences by drawing from four strands of the statistics literature that study the choice of graphical parameters (e.g. Calonico et al., 2015, Li, Munk, Sieling, and Walther, 2020), their effects on visual inference (e.g. Cleveland and McGill, 1984a), the evaluation of visual inference through comparison with econometric inference (e.g. Majumder et al., 2013), and the ocular process underlying visual inference via eyetracking technology (e.g. Hofmann, Follett, Majumder, and Cook, 2012). To guide our study design and help interpret our empirical results, we construct a general conceptual framework, which extends to future studies of visual inference in other contexts. ii) This is also

the first paper within economics that uses lab experiments to study empirical methods, a new paradigm that can be applied to other important areas (see Section 5 for more details). Using 11 DGPs calibrated to RD and RK microdata allows us to carry out a more comprehensive and empirically relevant evaluation of econometric inference procedures than econometric studies that typically rely on two or three DGPs. iii) This paper relates to the recent economics studies on methodological transparency and scientific communication (e.g. Angrist and Pischke, 2010; Andrews, Gentzkow, and Shapiro, 2020; Andrews and Shapiro, 2020). In particular, the bin width choice can be seen in the lens of Andrews et al. (2020) as providing more or less data for the reader to draw inference on the parameter of interest while still being easier to process than the underlying microdata. And iv), we add to the literature on expert judgments (e.g. Camerer and Johnson, 1997) and expert forecasts of research results (e.g. Sanders, Mitchell, and Chonaire, 2015). Our finding that experts only partly anticipate our experimental results underscores the value of empirically evaluating visual inference and providing evidence-based guidance on graphical methods.

The remainder of this paper is organized as follows. In Section 2, we present our conceptual framework and formalize the measure we use to judge visual inference. In Section 3, we discuss the design of our experiments and studies, including the graphical parameters we test, the specification of the data generating processes, the design of the online survey, and the setup of our eyetracking study. Section 4 presents the empirical results. Section 5 concludes.

2 Conceptual Framework

In this section, we propose a conceptual framework for evaluating visual inference to guide study design and help with the interpretation of empirical results. The framework addresses the central questions arising from features of visual inference which set it apart from standard econometric inference. In particular, we outline how to aggregate visual inference performance across subjects given that they may reach different conclusions even when viewing the same graph and delineate what meaningful parameter our aggregate measure corresponds to.

Although RD graphs may serve other purposes, we view their most important function as accurately conveying discontinuity existence/magnitude at the policy threshold. According to Lee and Lemieux (2010), other purposes of RD graphs include i) helping to assess regression specifications and ii) allowing for the inspection of discontinuities away from the policy cutoff. But ultimately these other functions are moti-

vated by inference on the discontinuity at the policy threshold: i) can be viewed as reconciling visual and econometric inferences thereof and ii) informs the reader, under implicit global homogeneity assumptions, whether to believe the existence of a discontinuity at the policy threshold.

We therefore focus on binary classifications of a graph and treat type I and type II errors as the main performance measures for visual inference. However, the conceptual framework easily generalizes to assessing stated discontinuity magnitudes (which we elicit from experts).[1] A person commits a type I error in RD visual inference if she classifies a continuous graph as having a discontinuity, and a type II error if she classifies a discontinuous graph as continuous.

We introduce the following elements of the conceptual framework for RDD (the RKD framework is analogous). First, the vector γ denotes a combination of graphical parameters: bin width, bin spacing, the use of polynomial fit lines, inclusion of a vertical line, axis scaling, etc. Each entry of γ represents the value of a particular graphical parameter. In our experiments, we randomly assign each participant to view only graphs generated with a certain fixed value of γ. Ideally, one could run a large experiment with a full factorial design to test all combinations of the graphical values outlined above, but resource constraints force us to test a subset of the graphical parameter space via a sequence of studies as described in Section 3.3.

The combination (g,d) denotes the probability model underlying an RD dataset. g encompasses four elements: i) the distribution of the running variable X; ii) the conditional expectation function $E[\tilde{Y}|X=x]$ which is continuous at $x=0$; iii) the distribution of the error term u where $\tilde{Y} \equiv E[\tilde{Y}|X=x]+u$; iv) the sample size N. We note that the variable Y can represent the outcome, baseline covariates, or treatment take-up, so this framework applies to all graphs typically included in RD studies, including those from a fuzzy design. Intuitively, g specifies everything in the probability model except for the discontinuity, including the shape of the conditional expectation function. The discontinuity then results from shifting the right arm of the smooth function $E[\tilde{Y}|X=x]$ by some discontinuity level d, that is, $Y = \tilde{Y} + d \cdot 1_{[X \geqslant 0]}$ (which implies that $E[Y|X=x] = E[\tilde{Y}|X=x] + d \cdot 1_{[X \geqslant 0]}$, and we provide a graphical illustration in Section 3.2). Typically, the (g,d) combination is jointly referred to as the "data generating process," but we separate out the discontinuity level d and call g the DGP for ease of exposition.

We think of each (bivariate) RD dataset as a realization from the probability model (g,d). Realizations are indexed by ε, which can be thought of as the seed in statistical software. Taken together, the triplet

[1] Theoretically, we could attempt to obtain the confidence set of visual inference by asking the respondents whether the discontinuity is equal to d for a range of possible values for d. Given the practical challenges in implementation, however, we leave this approach to future work.

(g,d,ε) uniquely determines an RD dataset and, in fixing the graphical parameters for said dataset, the quadruplet (γ,g,d,ε) determines an RD graph. We will refer to (γ,g,d), that is the set of all possible graphs of type γ resulting from a given probability model, as a "meta-graph."

When presented with the same RD graph, there are several reasons why participants may draw different visual inferences. Some participants may be more skilled than others at classifying a discontinuity because they have received more training in statistics, have more experience with RD graphs, or otherwise have superior ability. Participants may also apply different intrinsic "p-value" cutoffs, as some may be more likely than others to classify any given graph as having a discontinuity. We use ϕ to capture these human characteristics that affect graph perception, whether or not they are observed by us.

The probability that a participant with characteristics ϕ reports that a discontinuity exists in RD graph (γ,g,d,ε) is denoted by $\tilde{p}(\gamma,g,d,\varepsilon,\phi)$. From casual observation, we know that the same participant may be influenced by idiosyncratic elements not encapsulated in ϕ and classify the same graph differently on different days. The probability formulation $\tilde{p}$ allows these factors to affect visual inference.

We now define the type I and type II error probabilities which we use to gauge participant performance. First, averaging $\tilde{p}$ over realizations ε and participant characteristics ϕ leads to the quantity

$$p(\gamma,g,d) \equiv E_{\varepsilon,\phi}[\tilde{p}(\gamma,g,d,\varepsilon,\phi)].$$

This is the probability that a randomly chosen participant reports a discontinuity in the meta-graph (γ,g,d). A high value of p indicates a high classification error probability for DGP g when the true discontinuity d is zero (type I error), but a low classification error probability when d is nonzero (type II error). Formally, the DGP-specific or g-specific type I and type II error probabilities for graphical parameter γ are defined as:

$$g\text{-specific type I error probability: } p(\gamma,g,0)$$

$$g\text{-specific type II error probability: } 1-p(\gamma,g,d) \text{ for } d \neq 0.$$

Conceptually, we can further average $p(\gamma,g,d)$ over the space of DGPs $\mathscr{G}$ (discussion of $\mathscr{G}$ after Assumption 1 below) to arrive at the *overall* discontinuity classification probability for γ:

$$\bar{p}(\gamma,d) \equiv E_{g\in\mathscr{G}}[p(\gamma,g,d)].$$

Correspondingly, the overall type I and type II error probabilities are defined as

$$\text{overall type I error probability: } \bar{p}(\gamma, 0)$$

$$\text{overall type II error probability: } 1 - \bar{p}(\gamma, d) \text{ for } d \neq 0.$$

In this paper, we design experiments to estimate the type I and type II error probabilities as defined above. For each meta-graph (γ, g, d), we randomly sample M participants and realizations. Participant i is shown RD graph $(\gamma, g, d, \varepsilon_i)$, where i and ε_i both take on values in the set $\{1, ..., M\}$, and is asked to assess the presence of a discontinuity. Let the binary variable $R_i(\gamma, g, d, \varepsilon_i)$ denote participant i's discontinuity classification, which equals one if the participant reports a discontinuity at the policy threshold. If the M participants and realizations come from simple random sampling from the distributions of ϕ and ε, then the following assumption holds:

Assumption 1. *For a given DGP g, the $R_i(\gamma, g, d, \varepsilon_i)$'s are i.i.d. with $E[R_i(\gamma, g, d, \varepsilon_i)] = p(\gamma, g, d)$.*

A natural estimator for $p(\gamma, g, d)$ is the sample average of discontinuity classifications:

$$\hat{p}(\gamma, g, d) = \frac{1}{M} \sum_i R_i(\gamma, g, d, \varepsilon_i).$$

In Proposition 1 in Appendix A.1, we state the distribution of $\hat{p}(\gamma, g, d)$ and show the estimator to be unbiased and consistent as $M \to \infty$ for $p(\gamma, g, d)$ under Assumption 1.

To estimate the overall probability $\bar{p}(\gamma, d)$, we need to sample from the DGP space $\mathscr{G}$. While it is difficult to formally define $\mathscr{G}$, we think of the data used in empirical RD research as realizations when sampling from $\mathscr{G}$. To that end, we can specify J DGPs that approximate data from existing research and present graphs generated with discontinuity d for each DGP g_j $(j = 1, ..., J)$ to a distinct group of M participants for a total of $M \cdot J$ participants and visual discontinuity classifications.

Assumption 2. *The DGP g_j's are randomly sampled from $\mathscr{G}$.*

A natural estimator for $\bar{p}(\gamma, d)$ is

$$\hat{\bar{p}}(\gamma, d) \equiv \frac{1}{J} \sum_j \hat{p}(\gamma, g_j, d) = \frac{1}{M \cdot J} \sum_{i,j} R_i(\gamma, g_j, d, \varepsilon_i),$$

the average of discontinuity classifications across the $M \cdot J$ classifications. In Proposition 2 in Appendix A.1, we state the distribution of $\hat{\bar{p}}(\gamma, d)$ and show the estimator to be unbiased and consistent (as $J \to \infty$) for $\bar{p}(\gamma, d)$ under Assumptions 1 and 2 (given that $J = 11$ in our experiments, consistency here is a conceptual statement implying that were we to incorporate DGPs from more RD studies, our type I and type II error

rates would be closer in probability to the population parameters of interest).

For brevity, we refer to the type I error probabilities $p(\gamma, g, 0)$ and $\bar{p}(\gamma, 0)$ as the DGP-specific and overall sizes of visual inference respectively, and refer to one minus the type II error probabilities—$p(\gamma, g, d)$ and $\bar{p}(\gamma, d)$ for $d \neq 0$—as the respective powers. This nomenclature is consistent with conventions in the statistics literature except for calling the overall type I error probabilities $\bar{p}(\gamma, 0)$ size. Conventionally, size is defined as the *supremum* of the type I error probability over the set of DGPs, whereas $\bar{p}(\gamma, 0)$ is an *average*. We choose not to adopt the conventional size definition for two reasons. First, finding the DGP that results in the highest type I error probability for visual inference is difficult. Second, that DGP may be very different from typical DGPs that underlie actual RDD datasets, which suggests that adopting the conventional size definition may be overly conservative.

We can also define the type I and type II error probabilities of the inference procedure based on a given econometric estimator, $\hat{\theta}$, but with two adjustments. First, γ is no longer an argument in these expressions because we directly apply $\hat{\theta}$ to the dataset (g, d, ε). Second, we need to specify the level of the testing procedure based on $\hat{\theta}$, which we set to 5%, the prevailing standard in empirical studies. Because the definitions of these probabilities and their estimators are similar to the quantities defined above, we omit them here.

In subsequent sections, we empirically trace out $\hat{\bar{p}}(\gamma, d)$ as functions of d, which are *power functions* that concisely summarize the type I and type II error probabilities. We evaluate visual inference by comparing its power functions $\hat{\bar{p}}(\gamma, d)$ across γ, between expert and non-expert populations, and against the corresponding power functions for various econometric inference procedures. We discuss the implementation of the inference procedure on the differences between the visual and econometric inference functions when we present our empirical results in Section 4, as its details depend on the design of our experiments.

In summary, we use the type I and type II error rates to measure the performance of an inference method. The type I error rate for RD visual inference is the fraction of continuous graphs participants incorrectly classify as having a discontinuity, and the type II error rate is the fraction of discontinuous graphs incorrectly classified as being continuous. Our framework allows us to interpret these rates as unbiased and consistent estimates of the type I and type II error probabilities, which represent the type I and type II probabilities a randomly chosen person commits when classifying a graph generated from a representative DGP.

3 Description of Experiments and Studies

3.1 Graphical Parameters Tested

We test the effects of bin width, bin spacing, parametric fit lines, vertical lines at the policy threshold, and y-axis scaling. Additionally, we test two RK-specific parameters: detrending the data with a global linear regression prior to creating the graph and plotting estimated first derivatives instead of the levels of the data so that a kink appears as a discontinuity. We discuss each of these treatments in detail below.

The most studied graphical parameter in RD is the width of each bin in the binned scatter plot. The first class of bin width selection algorithms comes from Lee and Lemieux (2010): start with some number of bins, double that number, test whether the additional bins fit the data significantly better, and repeat until the test fails to reject the null. Calonico et al. (2015) propose two bin width selection algorithms based on different econometric criteria. The first and more in line with the convention of the nonparametric regression literature is the bin selector that minimizes the IMSE of the bin-average estimator of the CEF, where the resulting number of bins increases with the sample size n at the rate $n^{1/3}$. For their second bin selector, the MV selector, Calonico et al. (2015) state that they "choose the number of bins so that the binned sample means have an asymptotic (integrated) variability approximately equal to the amount of variability of the raw data." The resulting number of bins increases with the sample size more quickly, at the rate $n/\log(n)^2$. The IMSE-optimal bin width therefore selects fewer bins, and hence has larger bin widths, than the MV algorithm. In addition to these two algorithms, Calonico et al. (2015) provide an interpretation for any given number of bins as the output of a weighted IMSE-optimal algorithm. We use this result later to extend the MV bin selection algorithm to plotting estimated derivatives in RKD. Applying the Lee and Lemieux (2010) algorithms to our datasets typically leads to bin numbers in between the IMSE and MV selectors that tend to be closer to those of the IMSE. We therefore restrict our analysis to the IMSE and MV bin selectors which we describe in more detail in Appendix A.2.

Although the prevailing approach is to adopt evenly spaced bins, with 98% of the more than 100 RDD and RKD studies we compile and review for current practices taking this approach, this method has drawbacks in that the resulting bins may contain vastly different numbers of observations, or even none at all. This can happen when the distribution of the running variable is far from uniform. As a remedy, Calonico et al. (2015) also propose quantile-spaced bins where each bin contains the same number of observations. Both spacings support IMSE and MV bin selectors, and we test each of these combinations.

Following suggestions by Imbens and Lemieux (2008) and Lee and Lemieux (2010), RD and RK graphs frequently feature parametric fit lines and a vertical line at the treatment threshold. Both papers suggest fit lines improve "visual clarity" by approximating the conditional expectation functions, and the default in the popular `rdplot` command by Calonico et al. (2015), for example, uses piecewise global quartic regressions on each side of the policy threshold. 10 of the 11 RDD papers on which we calibrate our DGPs include fit lines. For the six of these papers that generate fit lines using polynomial regressions, we use the same polynomial order as in the source graph, and in the remaining cases, our team unanimously decides on the fit that best matches the original fit line or the data (we could also use a formal data-driven approach to select the polynomial order, but different criteria from Lee and Lemieux (2010) lead to widely different recommendations as shown in Table A.1). The presence of a vertical line at the cutoff is designed to visually separate observations above and below that cutoff. We test all combinations of these two treatments except for including the fit line and not having a vertical line, which is used infrequently in practice (our literature review shows that fewer than 10% of papers use this combination).

The motivation for rescaling graph axes comes from Cleveland, Diaconis, and McGill (1982), who note that correlations on scatter plots seem stronger when scales are increased (see Figure A.1 for an example). We use two axis scaling options in our experiments. First is the default output returned by Stata 14. Second, we double that range by recording the range of the y-variable from the default graph, increasing the bounds by 50% of the original range in each direction, resulting in a graph where the data are condensed along the vertical axis. We do not manipulate the scale of the x-axis because our survey of the literature suggests that this is not a common adjustment. A related decision researchers encounter is the range of the running variable to use in producing graphs: should they use the entire dataset or only a subsample close to the policy threshold? We do not test this margin of adjustment in our experiments due to the difficulty in generalizing the findings from such an exercise. Suppose we find that selecting 50% of the observations closest to the threshold improves visual inference, should researchers "chop" the sample they are already planning to use? And after doing so, will it be beneficial to chop again? One could argue for testing the effect of using the full sample versus the subsample falling within the IK or CCT bandwidth (see Section 4.4.1 below for review), but these bandwidths themselves depend on the full sample—the first step in bandwidth calculation is a (semi-)global regression—and it is not even clear how we should define the "full" sample. In fact, some of the replication data used for our DGP calibration are already subsets of a larger sample.

There are other RD graphical parameters we do not test in our experiments. One is plotting confidence

bands around the binned averages or fit lines. However, confidence intervals are too complex to explain to the non-experts in our short tutorial without potentially affecting the way participants think about the classification task itself, and therefore we do not experimentally test their effects on visual inference. Because the moderate size of our expert pool makes it unsuited to randomized experiments, we refrain from testing other graphical parameters.

The first of our RK-specific treatments is detrending the data. This is simply running a global linear regression and plotting the residuals instead of the original data, as shown in Figure A.2. Plotting these residuals may make kinks more visually apparent and is adopted by Lee et al. (2019).[2]

We test one new treatment that has not previously been used in practice: estimating and plotting local first derivatives rather than the levels of the data. We follow Cattaneo et al. (2020), who examine partitioning-based nonparametric regression and derive an IMSE-optimal number of partitioning knots to estimate the level or derivatives of a CEF. As in conventional RD and RK plots, the knots are determined separately on each side of the policy threshold with no overlap. Within each bin defined by these knots, we run a linear regression, and for each bin, we plot the corresponding linear coefficient. A kink in levels then appears as a discontinuity in first derivatives, as in Figure A.3. Because the MV bin selector does not apply to this setting, we instead extend the results from Cattaneo et al. (2020) to approximate a small-width bin selector. For each of our RD datasets, we compute the implied weights that justify the MV bin selector as weighted-IMSE-optimal, take the median value, and use that weight for a weighted IMSE calculation of how many bins to use for the small bins treatment when plotting estimated derivatives for RK.

Because the RKD DGPs we use are based on large administrative datasets, as is typical for RKD, both bin selectors will generally select many more bins than in the RD setting. Using equally spaced MV bins tends to result in many near-empty bins in the sparse tails of our datasets, as in Figure A.4, which likely hinders kink perception at the policy threshold. Because such graphs never appear in papers implementing RKD, we use quantile spacing for all kink graphs.

3.2 Creation of Simulated Datasets and Graphs

For both RDD and RKD, we specify data generating processes based on the actual data used in published research. For the RDD components of our study, we randomly sample 11 from a set of empirical RD papers published in the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal*

[2]In principle, one could detrend by residualizing against any smooth function; we use linear detrending for simplicity.

of Business and Economic Statistics, Journal of Political Economy, Quarterly Journal of Economics, Review of Economic Studies, and *Review of Economics and Statistics* that have replication data available to create our DGPs. We refer to these DGPs as RD DGP1-DGP11.

The calibration of each DGP g entails the specification of its four components: the distribution of the running variable X, the continuous conditional expectation function $E[\tilde{Y}|X = x]$, the distribution of the error term u, and the sample size N. Out of the 11 studies, two plot residualized outcomes to account for covariates, conceptually consistent with the covariate-adjusted RD estimation by Calonico, Cattaneo, Farrell, and Titiunik (2019) and the ideas of Angrist and Rokkanen (2015). We use the empirical distribution of the running variable from each of the 11 papers but normalize it to lie in $[-1,1]$ with 0 representing the treatment cutoff. We also remove the most extreme observations where $|X| > 0.99$, following Imbens and Kalyanaraman (2012) and Calonico et al. (2014). For two papers which feature semi-discrete running variables, we add small amounts of normal noise to the running variable to match the regularity conditions from Calonico et al. (2015). To create a continuous CEF, we fit global piecewise quintics (still following Imbens and Kalyanaraman, 2012 and Calonico et al., 2014) and vertically shift the right arm. We specify the distribution of the error term u as i.i.d. normal with mean zero and standard deviation σ, which we set as the root mean squared error of the piecewise quintic regression. We use the same number of observations as the original paper minus any observations removed while trimming the data. Plots of the resulting CEFs before we vertically shift their right arm to make them continuous are in Figure A.5, and Figure A.6 illustrates the construction process. We leave the full details of DGP specification to Appendix B.

Because the outcomes from the 11 papers are measured in different units, we specify discontinuity levels d as multiples of σ. d takes on 11 values: $0, \pm 0.1944\sigma, \pm 0.324\sigma, \pm 0.54\sigma, \pm 0.9\sigma, \pm 1.5\sigma$. We include $d = 0$ in order to measure type I error rates. We choose the upper bound $|d| = 1.5\sigma$ based on our own visual judgment: it represents the point at which we expect every reasonable person to say a graph from any of our 11 DGPs features a discontinuity. The nonzero magnitudes of d are equally spaced on the log scale. We use a log rather than linear scale to generate more graphs with smaller discontinuities, which are harder to detect, in order to better capture the shape of the power functions.

Our discontinuity magnitudes are similarly distributed to those observed in the main outcome graphs in our literature review. The average absolute value of the t-statistics in our datasets from piecewise quintic regressions is 5.0 with a standard deviation of 5.3, compared to the observed mean of 3.9 with a standard deviation of 6.4. If we instead compare the distributions of the absolute value of the discontinuity divided

by the control magnitude (the left intercept of the CEF), the means are similar, 1.3 in our datasets and 1.9 in the field, while our standard deviations are somewhat smaller at 2.2 compared to 7.7.

In the RKD case, most studies use restricted administrative datasets, resulting in a lack of publicly available replication microdata. Therefore, we fit 11 DGPs based on three studies: Card, Lee, and Pei (2009), Card et al. (2015a), and Card, Lee, Pei, and Weber (2015c). The DGPs are fitted on seven subsamples: the reduced-form sample of Card et al. (2009), the pre- and post-recession reduced-form samples of Card et al. (2015a), and the top and bottom kink first-stage and reduced-form samples of Card et al. (2015c). Although the latter two papers use restricted data from Austria and Missouri respectively, a member of our team has specified the DGPs from past work on these projects.

The process for calibrating RK DGPs is similar to that for RD. Data confidentiality creates a complication in that the actual data cannot be shared among team members, so instead of using the X variable from the actual data, the running variable used to create each graph is sampled from an estimated kernel density. For each of the seven samples, we obtain a CEF from piecewise quintic regressions like for their RD counterparts, except here we impose continuity as implied by the continuous treatment rule and the economic model in Card et al. (2015c). In addition, for each of the four subsamples of Card et al. (2015c), we follow the Monte Carlo simulations of Card, Lee, Pei, and Weber (2015d) and Card et al. (2017) and specify an additional CEF from a *global* quintic regression. We create these four DGPs that are smooth at the policy threshold out of concern that using only the other seven may be too favorable to visual inference based on the results of an RKD pilot (the pilot revealed both very low type I error rates and high power in a setting that did not require participants to distinguish between curvature and kinks at the policy threshold). We plot all resulting CEFs in Figure A.7. The CEFs for RK DGPs 5, 6, 9, and 10 are the global quintics, while the rest are piecewise quintics. We create CEFs with different kinks by rotating the right arm of the unkinked CEF, and we specify the kink magnitude as multiples of ξ, which denotes in radians the asymptotic standard deviation of the angle estimator between the left and right arms of the CEF from a piecewise quintic regression. We use nine different kink levels: $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$. Similar to our RD DGPs, we also use a normal distribution for the error term u, but the specification of its standard deviation is more complex. We discuss in detail the calibration of our RK DGPs in Appendix B.

As argued in Section 2, we want our DGPs to be representative. Although we select the papers randomly, we also need to evaluate how well our DGPs approximate the actual data. To do this, we adapt the lineup protocol from Buja et al. (2009) and Majumder et al. (2013), which uses visual inference to conduct

hypothesis testing.[3] In our case, we test the null hypothesis that the original datasets come from the calibrated DGPs. Specifically, we present one graph of the original data randomly placed among 19 graphs from datasets drawn from the corresponding DGP. The goal is to identify the true dataset by choosing the graph that least resembles the others. If the viewer does not select the original graph, then we cannot reject the null hypothesis. The probability of identifying the graph produced from the original data (or the type I error probability) among the 19 simulated datasets is 5% $(1/20)$ for a single reader. For our lineup protocol, each graph is a binned scatter plot using the MV bin selector. An example of the combined plots is in Figure A.8, with the coordinates of the true graph in this footnote.[4]

Based on visual testing among the authors, we cannot identify the graph from the true data for eight out of our 11 RD DGPs, which supports the idea that our DGPs approximate the original datasets well. For three DGPs, however, there is an obvious difference, such as in Figure A.9. All three "fail" seemingly because of the misspecification of the variance structure of the error term u. Recall that we specify u as being i.i.d. across observations and homoskedastic. But in Figure A.9, for example, the running variable is time, and there is positive serial correlation in the outcome. As a consequence, the outcome variability in the binned scatter plot is understated when u is assumed to be i.i.d. Nevertheless, we adhere to the i.i.d. specification because it is standard in Monte Carlo exercises to evaluate RD estimators and makes it straightforward to implement the uniformly most powerful test in Section 4.

Another caveat of our DGP specification is that using quintic regressions can lead to overfitting. Some of our graphs feature high variation across bins in the tails due to the sensitivity of the fitted piecewise quintics. We follow the practice of Imbens and Kalyanaraman (2012) and Calonico et al. (2014) in our DGP specification while acknowledging this potential drawback of using quintics. These caveats notwithstanding, the lineup protocol offers a novel and transparent method to evaluate DGP specifications, and it is because of its usefulness that we can identify the aforementioned limitations at all.

3.3 Non-Expert Experiments

In our randomized experiments, we present non-expert participants with binned scatter plots made from our DGPs and ask them to classify the graphs as having a discontinuity/kink or not. We conduct five phases

[3] Because of confidentiality concerns with the RK microdata from Card et al. (2015a) and Card et al. (2015c), we do not present lineups for the RK DGPs.

[4] The graph produced from data used in the original paper is in row $-3 \cdot 2 + 7$ and column $\sqrt{4+5} - 2$. The remaining graphs are generated from our specified DGP.

of computer-based experiments online through Cornell University Johnson College's Business Simulation Lab. Our subject pool consists of current and former Cornell students, Cornell staff, and non-student local residents with an expressed interest in focus groups or surveys. Although these educated laypeople are not the primary audience for academic research, RD graphs are sufficiently transparent that they are featured in popular media articles in publications such as *The New York Times*, *The Washington Post*, and *The Atlantic* (Dynarski, 2014; Sides, 2015; Rosen, 2015), suggesting the participants in our sample should be capable of interpreting the graphs.

Before the experiment, participants watch a video tutorial explaining how the graphs are constructed.[5] We do not instruct participants on how to make their decisions, e.g. whether only to look at points near the cutoff or mentally to trace out the CEF. The video contains an attention check with a corresponding question later in the experiment to ensure that subjects pay attention to the instructions. After the video, participants complete a series of interactive example tasks and receive feedback on their answers. As part of the instructions, we tell the participants explicitly that all, some, or none of the 11 graphs they classify may feature a discontinuity/kink.

In each phase of the experiment, we present participants with a series of RD or RK graphs using data generated as described in Section 3.2. In the RD experiments, participants see two graphs with zero discontinuities, one each of $\pm 0.1944\sigma, \pm 0.324\sigma, \pm 0.54\sigma, \pm 0.9\sigma$, and one of either 1.5σ or -1.5σ. Participants see one graph from all 11 DGPs in a randomized order. We have up to 88 participants per treatment arm, and every graph we generate is seen by only one participant. The setup is similar for RK, except we present each participant with 11 graphs plotting levels of the data and another 11 graphs plotting estimated first derivatives, for a total of 22 classification tasks. For each graph, we ask participants whether they believe there is a discontinuity (or kink for RK) at $x = 0$. To stimulate participant engagement and elicit participants' confidence in their response, participants can choose a bonus that is either based on a monetary wager which pays 40 cents if their judgment is correct but nothing otherwise, or a fixed payment of 20 cents irrespective of their performance. Under this bonus scheme, risk- or loss-averse participants will choose the fixed payment whenever their subjective probability of being correct is close to 50% (see Appendix C.1.1 for details).

Because running an experiment with $2^5 = 32$ RDD treatment arms is infeasible with our resources, we conduct our experiment in phases, testing only a few treatments in each phase. To inform best practices in graphical representation, we are mainly interested in the direction of the effects of graphical parameters, e.g.

[5]The video tutorial is available at `https://www.dropbox.com/s/3q5bj0adxis4vf2/RD_video_tutorial.mp4?dl=0`.

whether using small bins increases or decreases type I error rates relative to using large bins. Therefore, we assume that our graphical parameters do not have "qualitative interactions" (see Gail and Simon, 1985), i.e., the direction of the effect of one parameter does not change with other parameters. We can test this assumption on the two treatments that we interact in each phase, and it appears to hold.

Tables A.2 and A.3 list the graphical parameters we test and hold fixed for various experimental phases. In Phase RDD1, we test both bin width and axis scaling options. In Phase RDD2, we test bin widths and bin spacings. In Phase RDD3, we test imposing fit lines and a vertical line at the treatment threshold. In Phase RKD, we test combinations of bin widths and detrending the data and separately test bin widths for graphs of local derivatives. Because RKD participants see graphs both with levels of the data and estimated first derivatives, we have a between-subjects design for the RK graphical treatments of bin width and detrending, but a within-subject design for plotting levels and derivatives.

Participants receive a base pay of $3 for being in the experiment and a bonus of 0, 20, or 40 cents per classification depending on the participant's wager selection and whether they are correct. We do not give participants real-time feedback but do report earnings and the total tally of correct classifications after a short exit survey soliciting demographic information and comments at the end of the experiment. The study takes participants approximately 15 minutes to complete.[6]

3.4 Expert Study

In addition to our non-expert experiment, we conduct a study with researchers in economics and related fields who work on topics that often employ RDDs and RKDs. We collect data at three technical social science seminars and online by contacting randomly selected members of the National Bureau of Economic Research in applied microeconomic fields (aging, children, development, education, health, health care, industrial organization, labor, and public) and Institute of Labor Economics (IZA) fellows and affiliates. After removing a total of six responses from participants who completed the survey more than once, did not provide a valid email address for payment, or were not part of our recruited sample, we are left with 143 expert responses for RD and 48 for RK.

This expert study allows us to answer two questions. First, how do classification accuracy and the impacts of graphical techniques differ between experts and non-experts? And second, can experts correctly

[6]The experiments are programmed in oTree (Chen, Schonger, and Wickens, 2016) and pre-registered at the AEA RCT registry (Korting et al., 2019a) and the Center for Open Science's OSF platform (Korting et al., 2019b).

predict which graphing options perform best for our non-expert sample? This second question speaks to experts' ability to predict which visualization choices are most suited for interpretation by a lay audience. Because the success of a graphical technique ultimately lies in the reader's correct perception of graphs using it, it is important to understand whether experts' intuition regarding the relative advantages and drawbacks of alternative representation choices aligns with the evidence we find in practice. In related work on experts' ability to predict non-expert performance, DellaVigna and Pope (2018) find that economic experts are better than non-experts at estimating the effect of alternative incentive schemes on performance in a real effort task, but perform similarly to non-experts in terms of a simple ranking of incentive schemes.

Our expert study consists of two parts. The first is similar in structure to the non-expert experiment. Participants see a series of RD or RK graphs and are asked to classify them as (not) having a discontinuity/kink. To assess the accuracy of point estimates in addition to binary classifications of discontinuities, we also ask participants in the RD module for an estimate of the discontinuity magnitude whenever they report a discontinuity. Due to sample size limitations, we do not randomize graphical treatments in the expert study, and all participants see graphs with no fit lines, default axis scaling, and a vertical line at the treatment threshold. All expert RD graphs use equally spaced small bins, except for one seminar where participants see large bins, and all expert RK graphs use large bins with quantile spacing. Four randomly selected participants receive a base payment of \$450 plus a bonus payment of \$50 per correct discontinuity/kink classification. The bonus payment in the RD case does not depend on the accuracy of the magnitude estimate.

The second part of the expert study asks about experts' preferences and their beliefs regarding non-expert performance across alternative graphical parameters. We present experts with three discontinuity or kink magnitudes: $0, 0.54\sigma, 1.5\sigma$ for RD and $0, 0.324\xi, 0.9\xi$ for RK. For each RD magnitude, we present four graphs using the same underlying data, one for each combination of bin width and fit lines. We show graphs from the DGP where visual inference performs most closely to the average over all DGPs. For RK, we show only two graphs with both bin width options and without fit lines. We randomize graph treatment order between participants. At each magnitude, we ask the experts to indicate which of the four (two) treatment options they prefer and which of the four (two) treatment options they believe perform best and worst in our non-expert sample. To evaluate the experts' predictions about non-expert performance, we run a supplemental phase RDD4 in which we test all four treatments simultaneously across subjects. This is not necessary for RK, where we test all treatments in one phase.

3.5 Eyetracking Study

A followup to our observed performance differences across individuals is to ask why some people are better at classifying discontinuities than others and whether visual patterns can predict performance. To answer these questions, we conduct an eyetracking study through Cornell University's Mobile Communication Lab. Eyetrackers provide insights into cognitive processes by tracing participants' eye patterns as they navigate visual information. These eye patterns are typically differentiated into *saccades* (rapid eye movements) and *fixations* (maintaining visual gaze on a single location). Fixations are believed to be a valid indicator for attention during complex information processing tasks such as reading (Just and Carpenter, 1980; Rayner, 1998) and have previously been used to study how physicians analyze electrocardiograms (Bond et al., 2014), how children "read" bar graphs (Garcia Moreno-Esteva et al., 2017), and how alternative presentations of health messages online improve recall (Bol et al., 2016).

The structure of the eyetracking study is similar to that of our non-expert experiment, and both experts and non-experts participate in the study. We recruit non-expert participants from the same Business Simulation Lab subject pool. All non-expert subjects receive the same instructions and training as in the online experiment before completing the classification task at a physical eyetracking terminal (see Appendix C.3 for details). We recruit experts from a pool of experienced graduate students and faculty at Cornell University. Expert and non-expert participants receive participation fees of $10 and $5, respectively, plus 40 cents for each correctly classified graph. In total, we recruit 11 non-expert and six expert subjects. We drop one non-expert participant from the sample because of the poor quality of their eyetracking data resulting from a low calibration score.

The eyetracking approach allows us to investigate the extent to which visual patterns predict participants' correct classification of a graph. Similar to the RD phases of our primary experiment, we present subjects with a series of 11 RD graphs, one for each DGP, but this time showing the same graphs in the same order to all participants. We keep the set of graphs fixed because the visual patterns that lead to correct discontinuity classification may vary from graph to graph, and due to the risk of visual fatigue affecting eye movements over time, we also keep the same graph order between subjects.

4 Results

4.1 RD Non-Expert Experiment Results

For each combination of graphical parameters in all phases of the regression discontinuity experiment, we compute power functions based on participants' classifications of graphs as having or not having a discontinuity. Using notation from Section 2, a DGP-specific power function represents the estimates $\hat{p}(\gamma, g, d)$ for graphical parameters γ and DGP g across different levels of discontinuity d. An overall power function represents the estimates $\hat{\bar{p}}(\gamma, d)$.

The intercept of a power function indicates the type I error rate or size as defined in Section 2. At all other discontinuity magnitudes, the power function represents the proportion of graphs with discontinuities that participants classify correctly, which can be interpreted as the statistical power in detecting the discontinuity when the DGP is chosen uniformly randomly from the 11 possibilities. The ideal inference method has a small intercept before quickly rising to high power. Plots of the overall and DGP-specific power functions for each phase are in Figures 1 and A.10 respectively. The x-axis in these graphs is the magnitude of the discontinuity divided by the DGP-specific σ. This normalization facilitates aggregation and comparisons across DGPs, which then have six identical discontinuity magnitudes: 0, 0.1944, 0.324, 0.54, 0.9, and 1.5. Estimated treatment effects are in Tables A.4-A.7. Because we effectively adopt stratified randomization in the design of our experiments as described in Section 3.3, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude, we obtain these estimates by regressing the participants' responses on treatment indicators and stratum fixed effects.

Phase RDD1 tests the four combinations of the bin width treatments (large IMSE-optimal bins and small MV bins) with the y-axis scaling treatments (the default in Stata 14 and double that range). Comparing power functions, the large bins have significantly higher type I error rates relative to the small bins. While both small bin treatments lead to a size of approximately 5%, the large bins have a size of around 20% to 25%. The large bins have a power advantage over the small bins, but the treatments converge in power as the discontinuity magnitude increases. In contrast to bin widths, axis scaling has little effect on participant perception. Based on this result, we use Stata's default scaling for all subsequent phases.

In Phase RDD2, we again test the two bin width treatments, this time interacted with the two bin-spacing treatments: even spacing and quantile spacing. Note that large bins and small bins with even spacing appear in both Phases RDD1 and RDD2. With this design, we can gauge the stability of visual inference

across different samples from the non-expert population, and it is encouraging to see the results for these two treatments being virtually identical across phases. Comparing the power functions for these repeated treatments with their new quantile-spaced versions, we see that the evenly spaced and quantile-spaced power functions are very similar. We conclude that bin spacing has a small or null effect on visual inference for the DGPs we test.

Phase RDD3 tests three treatments: the inclusion of a vertical line at the treatment threshold with and without polynomial fit lines and the omission of both the vertical line and fit lines. We find that the vertical line at the cutoff makes little difference in perception. Fit lines, on the other hand, appear to increase both size and power in this phase, in line with a common concern that they may be overly suggestive of discontinuities.

Jointly, these three phases of experiments suggest that the presence of fit lines and the bin width choice have the largest impact on visual perceptions of discontinuities. We therefore base our analysis of expert preferences and expert predictions about non-expert performance on the interaction of these two treatments and run a final phase of experiments, Phase RDD4, directly comparing the four possible treatment combinations. Interestingly, while the effects of bin width choice are once again robust across phases, the effects of fit lines are more muted in this phase. In particular, the treatment with small bin and fit lines has a size of only 0.052 in Phase RDD4 but 0.175 in Phase RDD3. This finding suggests that we cannot conclude that fit lines unequivocally result in an increase in type I errors.

We provide additional results in Appendices D and E. Appendix D includes evidence for the balance of covariates across treatment arms and the measurement of predictive power of demographic and DGP characteristic variables on visual inference performance. Appendix E provides an answer to the question of how large the t-statistics need to be for readers to visually detect a discontinuity.

4.2 RK Non-Expert Experiment Results

We conduct one phase of the non-expert experiment focusing on the regression kink design. Here we test two separate sets of treatments. In the first set, we compare all combinations of IMSE and MV bin widths with detrending and not detrending the data. Power functions based on the raw or detrended data are plotted in Figure 2 with the treatment effect estimates in Table A.8. Type I error rates are universally high for kinks, around 15% to 20%. Breaking the DGPs into the seven piecewise quintics and the four smooth global quintics with peaks at $x = 0$ in Figure A.11 shows that the high type I error rate is driven by the smooth

DGPs. Small bins again decrease the type I error rate, but the effect is much smaller than for discontinuities at about one percentage point. Detrending the data prior to plotting has positive effects on power while maintaining a similar type I error rate.

The power functions from plotting the estimated first derivatives in the data and identifying a discontinuity in them are also shown in Figure 2 with corresponding treatment effect estimates in Table A.9. Results for this treatment are not compelling, with high rates of both type I and II errors. For the small bin treatment, the size is above 20% and the power peaks at around 30%, while the large bin treatment's size is almost 40% and the power does not reach 70%. These power functions are dominated by the levels treatments, and we advise against plotting estimated derivatives in RKD applications.

4.3 Expert Study Results

4.3.1 RD Expert Study Results

For the experts in our RD sample, we show most participants (95 out of 143) graphs with small bins, no fit lines, even spacing, default y-axis scaling, and a vertical line at the policy threshold, which have the lowest type I error rate in our non-expert experiment. We present the expert power functions and the non-expert power functions based on graphs generated with the same parameters in the left panel of Figure 3, and we plot the differences between the power functions in Figure A.12. This figure also includes pointwise 95% confidence intervals generated by applying the large sample approximation described at the end of Appendix A.1 and by assuming independence between the experts and non-experts. The two groups perform similarly, with experts having slightly worse size (approximately 8% to the non-expert 5%) and slightly greater power. The only statistically significant differences are for the experts' marginally superior powers at the 0.1944 and 0.324 discontinuities.

In addition to the aforementioned treatment, we show experts in one seminar pool (48 out of 143) graphs using the large bins and no fit lines treatment. We compare them with non-experts under the same treatment in the left panel of Figure 3. The two groups perform similarly. Experts again have a slightly worse type I error rate than non-experts here, and both are well above the rates for experts and non-experts with small bins. The expert and non-expert power functions are not statistically significantly different anywhere.

4.3.2 RK Study Expert Results

Experts in our RK sample see graphs with large bins, no fit lines, quantile spacing, and a vertical line at the policy threshold. In addition to using quantile-spaced bins as mentioned in Section 3.1, we use only large bins because on our kink datasets based on large administrative data, the small bin algorithm selects almost 1,000 times more bins than the large bin algorithm, resulting in graphs that do not look like those that commonly appear in RKD papers (for RD, the small bin algorithm only selects about 16 times more bins). We compare their performance with that of non-experts in the same treatment in the right panel of Figure 3 and Figure A.13. Although there is virtually no size difference, experts have substantially better power for moderately sized kinks.

4.3.3 Expert Preferences and Predicting Non-Expert Performance

For RD, we present experts' preferences and their beliefs about non-expert performance across the four considered treatments in Figure A.14. When asked about graphing options for the main graph of a paper that conveys the treatment effect, most experts report preferring small bins, usually with fit lines. These results hold at all three discontinuity magnitudes considered, including zero. Experts' predictions about the most effective treatments for non-experts tend to mirror their preferences. By a large margin, experts believe small bins with fit lines to be the most efficacious treatment for non-experts at all discontinuity magnitudes. Conversely, most experts view large bins without fit lines least favorably in the context of non-expert performance.

Comparing the expert predictions for RD to our experimental data from Phase RDD4, we find substantial discordance for the effects of bin width choice on non-expert classification accuracy. The best- and worst-performing treatments at each discontinuity magnitude have + and - signs, respectively, in Figure A.14. The actual power functions are shown in Figure 1, and the ones based only on the DGP used in the example graphs shown to experts are in Figure A.15. While a majority of experts correctly identifies the bin width treatment with the best size control (i.e. most experts prefer small bins at the zero discontinuity level, either with or without fit lines), there is also significant expert support for the large bin with fit lines treatment, even when there is no discontinuity, which exhibits the greatest type I error rate in our sample. In addition, experts fail to predict the type I vs type II error tradeoff presented by the bin width choice: most experts expect large bins to perform worst even at large discontinuities, while we find this treatment arm has the greatest power

in those cases. Although in the actual power functions, the effects of bin width are much more pronounced than the effect of fit lines, we find more expert disagreement regarding non-expert performance along this dimension. We also find expert predictions to be similar whether their own visual inference performance is above or below the median.

Figure A.16 shows experts' personal preferences and beliefs about non-expert performance for RK graphs with small bins and large bins, both without fit lines. For both themselves and the non-experts, experts clearly prefer large bins at every kink magnitude. Experts pick only one of the three best performers in the non-expert results in Figure 2, though differences between these treatments are generally smaller than in the RD setting, making prediction more difficult.

4.4 Visual vs Econometric Inference

4.4.1 RD Visual vs Econometric Inference

In this section, we compare the performances of visual inference from our small bin expert sample with econometric inference based on various RD estimators. For each estimator, we present graphs of both the overall power functions and the difference between them. For a fair comparison, we base the estimators' power calculations on their rejection decisions over the same set of datasets underlying the graphs seen by the experts. That is, the estimators "see" the same data as the experts, preventing differences driven by variation in sampling from the same DGP.[7]

As a benchmark, our first estimator comes from a correctly specified model: a global piecewise quintic regression with homoskedastic standard errors. The power function for a 5% test based on this estimator compared with human performance is presented in the left panel of Figure 4 with the differences between them in Figure A.17. The midpoint of each vertical bar in Figure A.17 represents the difference between point estimates from regressions of the inference decision on indicators for each discontinuity magnitude. We compute confidence intervals for these differences by assuming independence between visual and econometric inferences.[8] We cluster the visual inferences at the participant level and the econometric inferences at the DGP-seed level (there are 88 such combinations). The type I error rate is just under 8% for the ex-

[7] Out of the 11 RD studies, eight report main treatment effects and plot the data without controlling for covariates, and two report covariate-adjusted estimates and plot residualized outcomes. For these ten studies, visual and econometric inferences are based on the same data input, and our DGP specifications allow for an apples-to-apples comparison.

[8] As we show in Appendix F, we cannot reject the independence assumption at the zero discontinuity level. Visual inference tends to be positively associated with econometric inference at nonzero discontinuity levels, and thus the confidence intervals for the differences are conservative there.

perts and approximately 6.8% for the piecewise quintic, which is near 5% by design. The piecewise quintic estimator has higher power, though the difference is only significant at the 0.324 discontinuity.

Next, we implement the IK, CCT, and AK inference procedures, again plotting the corresponding power functions in the left panel of Figure 4 and their differences in Figure A.17. All three procedures build upon local linear regressions but take different approaches to conduct inference.

Strictly speaking, Imbens and Kalyanaraman (2012) do not study inference but propose an MSE-optimal bandwidth selector, the IK bandwidth:

$$\hat{h} = C \cdot \left(\frac{\hat{\sigma}^2(0^+) + \hat{\sigma}^2(0^-)}{\hat{f}(0) \cdot (\hat{\mu}^{(2)}(0^+) - \hat{\mu}^{(2)}(0^-))^2 + \hat{r}} \right)^{1/5} \cdot n^{-1/5}$$

where C is a constant determined by the kernel, $f(\cdot)$ is the density of the running variable, $\mu^{(2)}(0^{\pm})$ and $\sigma^2(0^{\pm})$ represent the second derivatives of the CEF and conditional variance on both sides of the threshold, r is a regularization term to prevent very large bandwidths, and the hats on the various quantities indicate that they are estimated. We refer to the "conventional" (terminology from Calonico et al., 2014) inference procedure practitioners typically implement in conjunction with the IK bandwidth as the "IK" inference procedure. Specifically, letting τ denote the true RD parameter and $\hat{\tau}$ the local linear estimator, we have

$$\sqrt{nh}(\hat{\tau} - \tau - h^2 B) \Rightarrow N(0, \Omega) \tag{1}$$

where B denotes the bias constant, which is a function of the second derivatives of the CEF on both sides of the threshold. Even though the normal distribution is centered around the constant $\tau + \sqrt{nh^5}B$ when h shrinks at the optimal rate $h = O(n^{-1/5})$ (as is the case for IK), "conventional" inference ignores the bias term $\sqrt{nh^5}B$. As seen in Figure 4, the 5%-level IK procedure is powerful, even more so than the piecewise quintic estimator, and is significantly better than visual inference at detecting discontinuities up to 1.5σ. But with this power comes a significant size distortion. When there is truly no discontinuity, the estimator still rejects the null hypothesis in 22.6% of datasets.

Third, we assess the performance of the inference procedure proposed by Calonico et al. (2014) (CCT) as implemented in the `rdrobust` Stata package. This procedure differs from IK and creates robust confidence intervals by estimating the bias B in expression (1) using another ("pilot") bandwidth b, creating a bias-corrected estimator $\hat{\tau}^{bc} = \hat{\tau} - h^2\hat{B}$, and accounting for the sampling variation in $\hat{B}$ under unconventional asymptotics where h/b converges to a positive constant. Calonico et al. (2014) also generalize IK and propose a new class of MSE-optimal bandwidth selectors, which we refer to as the CCT bandwidth(s). Note that although CCT's size is approximately 12.5%, as seen in the left panel of Figure 4, this could be a small

sample problem.[9] Because there are only eight draws from each DGP, i.e. ε takes only eight values for each g, we effectively have only 88 datasets, and all RD datasets within a DGP-seed pair are identical modulo the discontinuity level. In a separate Monte Carlo simulation with 1,000 draws, the size is in line with that of the experts at 7.0%. However, it is possible that the type I error rate of visual inference also decreases over graphs based on these alternative datasets—the realization of the disturbance term can impact both econometric and visual inference—and therefore we keep the comparison based on the 88 datasets. The CCT inference procedure is powerful, enjoying a significant 15 to 20 percentage point power advantage over expert visual inference at intermediate discontinuity levels.[10]

Finally, we apply the AK inference procedure from Armstrong and Kolesár (2018) and implemented in the R package `RDHonest`, which adapts Donoho (1994) and produces asymptotically valid and minimax optimal confidence intervals over the Taylor class of conditional expectation functions

$$\mathscr{F}_T(C;\chi_\pm) = \left\{ f_\pm : |f_\pm(x) - f'_\pm(0)x| \leqslant C_T|x|^2 \text{ for all } x \in \chi_\pm \right\}.$$

C_T is a researcher-supplied bound on the second derivatives on both sides of 0, which is an important tuning parameter and dictates the bias magnitude of a local linear estimator (hence, unlike CCT, AK's bias correction is nonrandom). When using the true second-derivative bounds, which requires omniscience about the DGPs, the AK inference procedure has a size of approximately 7%, and the power function is very close to that of the experts.[11] The package `RDHonest` also implements a rule-of-thumb curvature estimator as a default bound. This rule of thumb is obtained via a piecewise quartic regression, which approximates our piecewise quintic CEFs very well. Unsurprisingly, inference with these estimated bounds performs very similarly to inference using the true bounds, with a very small drop in size and a modest decrease in power. To further gauge how AK performs as a function of the tuning parameter, we use the maximum of the true

[9] While the CCT bandwidth remains the default and modal choice of `rdrobust`, new work by Calonico, Cattaneo, and Farrell (2020) proposes inference-optimal RD bandwidth selectors. As seen in Figure A.18, using these bandwidths reduces the size distortion relative to visual inference while retaining the power advantage.

[10] One may be concerned that the better size control of visual inference is a consequence of our experimental design where the majority (nine out of 11) of RD graphs feature a discontinuity. If subjects speculate that only about half the graphs feature a discontinuity, our size result is biased in favor of visual inference. As mentioned in Section 3.3 and Appendix C, we explicitly tell participants that "all, some or none of the 11 graphs you see in this survey may feature a discontinuity," and we present evidence in Appendix D.3 to further alleviate this concern.

[11] Armstrong and Kolesár (2020) propose analogous confidence intervals that maintain coverage and enjoy minimax optimality over a Hölder class of functions, which is determined by a global, as opposed to local, bound on the second derivative of the CEF. The corresponding power functions (not plotted) are similar when using the true global second-derivative bounds. Like Armstrong and Kolesár (2018, 2020), Imbens and Wager (2019) also adapt the idea of Donoho (1994). They propose an RD estimator through numerical optimization that is minimax mean-squared-error optimal over CEFs with a global second derivative bound. Because it performs similarly to Armstrong and Kolesár (2018) in simulations by Pei, Lee, Card, and Weber (2018) and can be computationally demanding, we do not implement the Imbens and Wager (2019) procedure here.

bounds across all 11 DGPs. These larger bounds result in a decrease in size of about four percentage points but a large drop in power, up to around 25 percentage points at intermediate magnitudes and 10 percentage points at the smallest and largest ones.

Inference procedures based on the IK and CCT estimators at the 5% level exhibit higher type I and lower type II error rates than visual inference. It is difficult to say whether these econometric inference procedures perform better or worse because there is no guidance on how to trade off type I and type II errors. We circumvent this problem by size-adjusting the inference procedures: we search for alternative critical t-values (or equivalently, critical p-values) such that the resulting type I error rate of the econometric inference procedure is equal to that of visual inference and then use those critical values to conduct inference. For IK, this critical t-value is 2.46, and it is 2.28 for CCT. We present the results for these size-adjusted inference procedures in the right panel of Figure 4. Despite the extent of their size differences relative to visual inference, both econometric procedures' powers only suffer by around 5-10 percentage points from this adjustment. When using these conservative critical values, the IK and CCT procedures retain their statistically significant power advantages over visual inference at moderate discontinuities while matching visual inference in size, as shown in Figure A.19.

In addition to asking experts to classify graphs as having or not having a discontinuity, we ask them to estimate the magnitude of the discontinuity. We take several steps to make human and econometric point estimates comparable. Because we ask participants to round their estimates to the nearest hundredth, we round estimators with the same precision. We similarly replace econometric estimates with 0 when the test fails to reject the null hypothesis.

Figure A.20 presents the root mean squared error (RMSE) of each estimator's point estimates by DGP, averaged over all discontinuity magnitudes. We do this by DGP to prevent scaling issues, as units for each DGP are different and results from one DGP are not directly comparable to those from another. Although there are a handful of DGPs where experts perform similarly to the econometric estimators, their point estimates generally have a greater RMSE than the estimators, and experts overall are worse at estimating magnitudes. We also decompose the human MSE gap relative to the piecewise quintic in Figure A.20 into its bias and variance components (see Appendix A.3 for details about this calculation). The DGPs are divided approximately evenly into those where bias or variance accounts for most of the MSE. In summary, although humans perform quite well at identifying the existence of discontinuities, their ability to estimate their magnitudes is not as strong.

4.4.2 RK Visual vs Econometric Inference

Paralleling the exercise for RD, we compare the performance of our expert pool with that of econometric inference procedures. We first try two 5% tests on the kink parameter using a correctly specified OLS regression model (piecewise quintic with homoskedastic standard errors) and the default CCT `rdrobust` procedure, which is based on a bias-corrected local quadratic estimator for kinks (Imbens and Kalyanaraman, 2012 and Armstrong and Kolesár, 2018 do not provide RK extensions of their RD methods). We compare the power functions of visual inference with those for the piecewise quintic estimator in Figures 5 and A.21. Despite being the correctly specified model, the piecewise quintic estimator has low power even at the largest kink level in our graphs. This is not surprising, as a kink of 0.9ξ corresponds to an asymptotic t-statistic of 0.9 for the piecewise quintic kink estimator, well below the 1.96 critical value used in the test. Results are similar for the default CCT `rdrobust` procedure, which has excellent size control but is highly conservative as seen in the same two figures.

The large divergence in the performances of visual and econometric RK inference raises the question of what the humans see that the two econometric procedures fail to capture. While this is a difficult problem, we can provide a partial answer by looking for econometric inference methods that better mimic visual inference. Beginning from the default CCT `rdrobust` procedure (bias-correction, local quadratic, triangular kernel, bandwidth regularization, and robust standard errors), we test alternative inference methods to find power functions that more closely resemble the experts'. We also plot the power functions of these additional inference procedures in Figure 5. Using a linear estimator instead of quadratic dramatically increases the power at larger kink magnitudes. Using conventional as opposed to robust standard errors and removing bandwidth regularization further increases power and brings the size closer to the experts'. The expert power function is sandwiched between those of conventional inferences based on local linear and local quadratic estimators with unregularized CCT bandwidths and uniform kernels.[12] We conclude that experts may be approximating a local linear or quadratic regression (or something in between) as a heuristic to classify graphs.

Although the econometric inference based on the correctly specified quintic regression lacks power on our simulated datasets, it is not the most powerful test we can conduct. Because of the normality of the error term in our DGP specification, we can implement a simple two-sided test, the uniformly most powerful

[12]For the local linear and quadratic estimators, the average regularized CCT bandwidths are 0.23 and 0.33, respectively. The corresponding unregularized bandwidths are 0.41 and 0.56. A bandwidth of 1 indicates a global regression.

unbiased test, for whether the kink is zero by imposing the true values of all parameters in the model except for the parameter of interest. The size of the test is approximately 8% on our sample and it has a power of one at all nonzero kink levels.[13] If we do not impose the true values of the other parameters (or nuisance parameters), the problem of computing the power function for the uniformly most powerful test becomes difficult (see recent progress by Elliott, Mäller, and Watson, 2015), and we leave it to future research.

To this point, we have presented the "marginal" power functions for visual and econometric inference procedures. In Appendix F, we extend this analysis to discuss the joint distribution of visual and econometric discontinuity/kink tests. Knowing the joint distribution i) informs us of their complementarity, i.e. whether visual and econometric inferences tend to be wrong on different data, and ii) is useful for combining inferences. We find that type II errors by experts are predictive of type II errors by various econometric inference methods, but the same is not true for type I errors. We conceptually illustrate how combining the binary visual inference decision with its econometric counterpart can refine overall inference.

4.5 Eyetracking Study Results

In our eyetracking study, we investigate whether certain eye patterns are predictive of correct discontinuity inference. Common practice in eyetracking is to divide the displayed image into "areas of interest" (AOIs) ex-ante and to measure the visual attention these AOIs receive. In a study by Byrne, Kalaji, and Niederdeppe (2018), for example, the AOIs include warning labels on images of cigarette packs, and the authors test whether the degree of visual attention therein predicts stated health risk beliefs and other outcomes ex-post. Our case is different: we are agnostic about the AOIs ex-ante, and instead search for eye patterns correlated with correct discontinuity inference.[14] If we identify robustly predictive eye patterns, we could then test in a follow-up experiment whether training participants based on these findings improves performance.

As explained in Appendix C.3, we choose the 11 graphs for this study by identifying those with the maximum difference in visual inference performance between non-experts and experts from the first three non-expert phases and two expert seminar samples. We begin by testing for differences in visual inference and attention between experts and non-experts in our eyetracking sample and present the results in Table

[13] We implement the same test for our RD DGPs as well, and its power function is practically the same.

[14] Because we show 11 different graphs, it is difficult to define consistent AOIs across the graphs. One might argue for using an optimal bandwidth to define AOIs. Because we are interested in the bandwidth for visual inference, we apply the `rdrobust` procedure to each of the binned datasets in the 11 graphs presented in the eyetracking study (the sample size of each dataset is the number of bins in the graph). The discontinuity classification based on the resulting 95% robust confidence intervals is only correct for six of the 11 graphs (55%), performing worse than our participants who classify 59% of graphs correctly on average.

A.10. Consistent with results from our early expert sample, the experts in the eyetracking study perform better—they correctly classify the existence of a discontinuity in 68% of the graphs, higher than the 53% rate for the non-experts. On average, the six experts spend more than twice as long on each graph and fixate on more than twice as many locations.

We then transform the eyegaze pixel coordinate data into the units of the (x, y) variables of each graph and compute a set of summary measures from the eyetracking microdata. These summary measures include the number of fixations, quantities that proxy for "visual bandwidth" (the horizontal range and horizontal standard deviation of the eyegaze coordinates), and the square root of the average squared deviation of the eyegaze coordinates from the true CEF scaled by σ (see Appendix G for details). We refer to the latter measure as the "eyegaze RIMSE" (root integrated mean squared error). We also fit RDD models to the eyegaze data using the CCT inference procedure and record the conventional and bias-corrected local linear estimates. We then compute the absolute deviation of these estimates from the true discontinuity in the DGP, which we also scale by σ.

The only measures that are significantly different between the experts and non-experts at the 5% level are time spent on each graph and whether the local RD eyegaze estimates are computable (which depends on the number of distinct eyegazes)—fewer RD estimates are computable from non-experts' eyegaze data. Significant at the 10% level, experts cover 30% more horizontal range (1.40 out of a maximum of 2).

Next, we assess the extent to which eyegaze patterns predict visual inference success. As mentioned above, the goal is to identify patterns that are causally linked to successful inference. To that end, our prediction exercise requires interpretability, and therefore we use the simple eyegaze summary measures in Table A.10 as inputs to various prediction models. In addition, we include the 11 graph identifiers as explanatory variables—eyegaze patterns differ across graphs, and graphs vary in their difficulty, so omitting the graph identifiers may lead us to erroneously conclude that certain eyegaze patterns are important. Because of the large number of covariates relative to the number of graphs, we fit penalized logit, pruned tree, bagging, and random forest models to prevent overfitting and we choose tuning parameters via five-fold cross validation where applicable.

The most important predictive variables appear to be the graph identifiers. Some of the graphs we test are easy for participants to classify while others are difficult, and knowing which graph a participant looks at is the most potent predictor. We do not detect any ocular patterns that robustly predict successful visual inference based on our small eyetracking study (the full details of the prediction exercise are in Appendix G).

We tentatively conclude that although experts outperform non-experts in this study, their eyegaze patterns are similar along most dimensions we consider, and the better performance is likely driven by the processing, as opposed to the acquisition, of visual signals.

5 Conclusion

This paper evaluates the performance of visual inference in RD and RK designs. Through a series of experiments and studies, we provide answers to questions in four related areas. First, how do graphical representation techniques affect visual inference? Second, do experts and non-experts differ in their visual inference performances, and are experts able to identify the most effective treatments for non-experts? Third, how does visual inference perform compared to common econometric inference procedures in RDD and RKD? And fourth, can we identify ocular patterns predictive of successful visual inference?

To answer the first question, we experimentally assess how five RD graphical parameters impact visual inference accuracy. We find that generating graphs with the Calonico et al. (2015) IMSE (large) bin selector leads to higher type I error rates but lower type II error rates relative to their MV (small) bin selector. Similarly, imposing fit lines leads to increased power but may come at the expense of size distortion. Bin spacing, a vertical line at the policy threshold, and y-axis scaling have little effect. For RKD, larger bin widths again lead to higher type I error rates, though to a smaller extent than in RDD, but generally lower type II error rates. Among the two additional RK parameters we test, detrending may improve power without distorting size, while plotting derivatives leads to worse size and worse power.

For the second set of questions, we recruit experts to participate in a similar study. Experts perform similarly to non-experts in reading RD graphs, with marginally lower type II error rates at moderate discontinuity levels. Experts' RK visual inference is considerably more powerful. When predicting different graphical representation techniques' effects on non-expert visual inference, experts only partly anticipate the tradeoff between type I and type II errors that arises.

Third, visual inference from RD graphs generated using the method with the smallest size achieves a lower type I error rate than econometric inference at the 5% level based on the Imbens and Kalyanaraman (2012) and Calonico et al. (2014) estimators, though the two econometric inference procedures offer considerable power advantages. The performance of visual inference is very similar to that based on the procedure suggested in Armstrong and Kolesár (2018). For RKD, type I error rates are higher for visual inference, but

only for DGPs with curvature that peaks near the threshold. Analyzing the joint distribution of visual and econometric tests reveals their complementarity: while they commit similar type II errors in RDD, there does not appear to be a strong association in their type I errors or in RKD. Systematically studying the settings in which visual inference outperforms econometric inference could be an important next step.

Finally, our eyetracking study does not reveal eyegaze patterns that robustly predict success in visual inference. Although the experts in this sample perform better than the non-experts, we cannot attribute this to their ocular movements, at least not in terms of summary measures such as their "visual bandwidth" or how well they trace out the underlying conditional expectation function.

A limitation of our study is the set of parameters we are able to experimentally test. As mentioned above, we do not impose fit lines with confidence intervals in our graphs, which researchers sometimes do, due to the difficulty in explaining it to non-expert participants. We also do not vary the size and color of the dots. However, these choices may impact inference due to their effect on visual attention and visual complexity as suggested by the literature on the psychological and neurological mechanisms underlying the processing of (visual) information (Hegarty, Canham, and Fabrikant, 2010; Kriz and Hegarty, 2007; Rosenholtz, Li, and Nakano, 2007; Wolfe and Horowitz, 2004). We leave these investigations to future work.

Another caveat for our results is that they are based on a specific set of DGPs. For example, while the bin spacing choice—equally spaced or quantile spaced—appears immaterial in our RD experiments, it could be important when the distribution of the running variable is farther from uniform than in our DGPs. On the other hand, the number of DGPs used in Monte Carlo simulations that lead to methodological recommendations is typically far lower than our 11, and those DGPs sometimes bear no semblance to real-world data. In addition, we test the validity of our simulated datasets by adapting the lineup protocol from Majumder et al. (2013) to assess the degree to which our DGPs approximate the original data.

Our study answers the call by Leek and Peng (2015) to provide empirical evidence on best practices in data analysis, and we hope it will spur similar investigations. One related topic to study follows the recent work by Cattaneo et al. (2019a), who, among other contributions, propose econometric tests of linearity and monotonicity based on binned scatter plots, which are motivated by studies such as Chetty et al. (2011) and Chetty et al. (2014). One could assess the impact of the graphical parameters on reader perception and compare visual and econometric linearity/monotonicity tests. Another closely related topic is structural breaks in time series econometrics, in which graphs serve practically the same purpose as those in RDD/RKD. Within time series econometrics, studying visual inference for unit root/stationarity analysis may also be promis-

ing.[15] In an influential textbook, Stock and Watson (2011) conduct an augmented Dickey-Fuller (ADF) test for the presence of a unit root in U.S. inflation. Upon finding that the test rejects a unit root at the 10% level but not at the 5% level, Stock and Watson (2011) write "The ADF statistics paint a rather ambiguous picture... Clearly, inflation in [the figure] exhibits long-run swings, consistent with the stochastic trend model." In this case, Stock and Watson (2011) apply visual unit root inference when the test statistic is marginal, which raises the question: can we leverage our eyes to begin with?

[15] In recent work, Shen and Wirjanto (2019) propose a new framework for stationarity tests, which formalizes the intuition that a visual characteristic of stationary time series is the infinite recurrence of "simple events" asymptotically.

References

AGRESTI, A. (1992): "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–153.

ANDREWS, I., M. GENTZKOW, AND J. M. SHAPIRO (2020): "Transparency in Structural Research," National Bureau of Economic Research Working Paper 26631.

ANDREWS, I. AND J. M. SHAPIRO (2020): "A Model of Scientific Communication," National Bureau of Economic Research Working Paper 26824.

ANGRIST, J. D. AND V. LAVY (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 144, 533–575.

ANGRIST, J. D. AND J.-S. PISCHKE (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of economic perspectives*, 24, 3–30.

ANGRIST, J. D. AND M. ROKKANEN (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344.

ARMSTRONG, T. AND M. KOLESÁR (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86(2), 655–683.

ARMSTRONG, T. B. AND M. KOLESÁR (2020): "Simple and Honest Confidence Intervals in Nonparametric Regression," *Quantitative Economics*, 11, 1–39.

BOL, N., J. C. VAN WEERT, E. F. LOOS, J. C. ROMANO BERGSTROM, S. BOLLE, AND E. M. SMETS (2016): "How Are Online Health Messages Processed? Using Eye Tracking to Predict Recall of Information in Younger and Older Adults," *Journal of health communication*, 21, 387–396.

BOND, R., T. ZHU, D. FINLAY, B. DREW, P. KLIGFIELD, D. GULDENRING, C. BREEN, A. GALLAGHER, M. J. DALY, AND G. D. CLIFFORD (2014): "Assessing Computerized Eye Tracking Technology for Gaining Insight into Expert Interpretation of the 12-Lead Electrocardiogram: an Objective Quantitative Approach," *Journal of electrocardiology*, 47, 895–906.

BUJA, A., D. COOK, H. HOFMANN, M. LAWRENCE, E.-K. LEE, D. F. SWAYNE, AND H. WICKHAM (2009): "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367, 4361–4383.

BYRNE, S., M. KALAJI, AND J. NIEDERDEPPE (2018): "Does Visual Attention to Graphic Warning Labels on Cigarette Packs Predict Key Outcomes Among Youth and Low-income Smokers?" *Tobacco Regulatory Science*, 4, 18–37.

CALONICO, S., M. D. CATTANEO, AND M. H. FARRELL (2020): "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs," *The Econometrics Journal*, 23, 192–210.

CALONICO, S., M. D. CATTANEO, M. H. FARRELL, AND R. TITIUNIK (2019): "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics*, 101, 442–451.

CALONICO, S., M. D. CATTANEO, AND R. TITIUNIK (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

——— (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110, 1753–1769.

CAMERER, C. F. AND E. J. JOHNSON (1997): "The Process-Performance Paradox in Expert Judgment: How can Experts Know so Much and Predict so Badly," *Research on judgment and decision making: Currents, connections, and controversies*, 342.

CARD, D., A. JOHNSTON, P. LEUNG, A. MAS, AND Z. PEI (2015a): "The Effect of Unemployment Benefits on the Duration of Unemployment Insurance Receipt: New Evidence from a Regression Kink Design in Missouri, 2003-2013," *American Economic Review: Papers & Proceedings*, 105, 126–130.

CARD, D., D. S. LEE, AND Z. PEI (2009): "Quasi-Experimental Identification and Estimation in the Regression Kink Design," Princeton University Industrial Relations Section Working Paper 553.

CARD, D., D. S. LEE, Z. PEI, AND A. WEBER (2015c): "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2453–2483.

——— (2015d): "Inference on Causal Effects in a Generalized Regression Kink Design," Upjohn Institute Working Paper 15-218.

——— (2017): "Regression Kink Design: Theory and Practice," in *Regression Discontinuity Designs: Theory and Applications*, ed. by M. D. Cattaneo and J. C. Escanciano, Emrald, vol. 38 of *Advances in Econometrics*, chap. 5, 341–382.

CATTANEO, M. D., R. K. CRUMP, M. H. FARRELL, AND Y. FENG (2019a): "On Binscatter," *arXiv preprint arXiv:1902.09608*.

CATTANEO, M. D., M. H. FARRELL, AND Y. FENG (2020): "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, 48, 1718–1741.

CHEN, D. L., M. SCHONGER, AND C. WICKENS (2016): "oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 126, 1593–1660.

CHETTY, R., N. HENDREN, P. KLINE, AND E. SAEZ (2014): "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *The Quarterly Journal of Economics*, 129, 1553–1623.

CLEVELAND, W. S., P. DIACONIS, AND R. MCGILL (1982): "Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased," *Science*, 216, 1138–1141.

CLEVELAND, W. S. AND R. MCGILL (1984a): "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, 531–554.

COHEN, J. (1960): "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37–46.

DAVENPORT, E. C. AND N. A. EL-SANHURRY (1991): "Phi/Phimax: Review and Synthesis," *Educational and Psychological Measurement*, 51, 821–828.

DELLAVIGNA, S. AND D. POPE (2018): "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126, 2410–2456.

DONOHO, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 238–270.

DYNARSKI, S. (2014): "What We Mean when We Say Student Debt is Bad," *New York Times*.

EELLS, W. C. (1926): "The Relative Merits of Circles and Bars for Representing Component Parts," *Journal of the American Statistical Association*, 21, 119–132.

ELLIOTT, G., U. K. MÄLLER, AND M. W. WATSON (2015): "Nearly Optimal t-Test when a Nuisance Parameter is Present under the Null Hypothesis," *Econometrica*, 83, 771–811.

GAIL, M. AND R. SIMON (1985): "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets," *Biometrics*, 41, 361–372.

GARCIA MORENO-ESTEVA, E., S. WHITE, J. WOOD, AND A. BLACK (2017): "Identifying Key Visual-Cognitive Processes in Students' Interpretation of Graph Representations Using Eye-Tracking Data and Math or Machine Learning Based Data Analysis," in *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1-5, 2017)*, Dublin City University.

HEGARTY, M., M. S. CANHAM, AND S. I. FABRIKANT (2010): "Thinking About the Weather: How Display Salience and Knowledge Affect Performance in a Graphic Inference Task." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 37.

HOFMANN, H., L. FOLLETT, M. MAJUMDER, AND D. COOK (2012): "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441–2448.

IMBENS, G. AND K. KALYANARAMAN (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of economic studies*, 79, 933–959.

IMBENS, G. AND S. WAGER (2019): "Optimized Regression Discontinuity Designs," *Review of Economics and Statistics*, 101, 264–278.

IMBENS, G. W. AND T. LEMIEUX (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635.

JUST, M. A. AND P. A. CARPENTER (1980): "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological review*, 87, 329.

KORTING, C., C. LIEBERMAN, J. MATSUDAIRA, Z. PEI, AND Y. SHEN (2019a): "A Study on Graphical Representation," *AEA RCT Registry (https://doi.org/10.1257/rct.4331-1.0)*.

——— (2019b): "A Study on Graphical Representation," *Retrieved from osf.io/jeax5*.

KRIZ, S. AND M. HEGARTY (2007): "Top-Down and Bottom-Up Influences on Learning from Animations," *International Journal of Human-Computer Studies*, 65, 911–930.

LEE, D. S. AND T. LEMIEUX (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

LEE, D. S., P. LEUNG, C. J. O'LEARY, Z. PEI, AND S. QUACH (2019): "Are Sufficient Statistics Necessary? Nonparametric Measurement of Deadweight Loss from Unemployment Insurance," National Bureau of Economic Research Working Paper 25574.

LEEK, J. T. AND R. D. PENG (2015): "Statistics: P Values are Just the Tip of the Iceberg," *Nature News*, 520, 612.

LI, H., A. MUNK, H. SIELING, AND G. WALTHER (2020): "The Essential Histogram," *Biometrika*, 107, 347–364.

MAJUMDER, M., H. HOFMANN, AND D. COOK (2013): "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of the American Statistical Association*, 108, 942–956.

NICKELL, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica: Journal of the Econometric Society*, 49(6), 1417–1426.

PEI, Z., D. S. LEE, D. CARD, AND A. WEBER (2018): "Local Polynomial Order in Regression Discontinuity Designs," Princeton University Industrial Relations Section Working Paper 622.

RAYNER, K. (1998): "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological bulletin*, 124, 372.

ROSEN, R. J. (2015): "Slight Changes in Yelp Ratings can Mean Huge Losses for Small Businesses," *The Atlantic*.

ROSENHOLTZ, R., Y. LI, AND L. NAKANO (2007): "Measuring Visual Clutter," *Journal of vision*, 7, 17–17.

SANDERS, M., F. MITCHELL, AND A. N. CHONAIRE (2015): "Just Common Sense? How Well Do Experts and Lay-People Do at Predicting the Findings of Behavioural Science Experiments," Harvard University Kennedy School Working Paper.

SHEN, Y. AND T. S. WIRJANTO (2019): "Stationarity as a Path Property," *Probability and Mathematical Statistics*, 39(2), 403–422.

SIDES, J. (2015): "How to Get Young People to Vote? Register Them Before they Turn 18." *The Washington Post*.

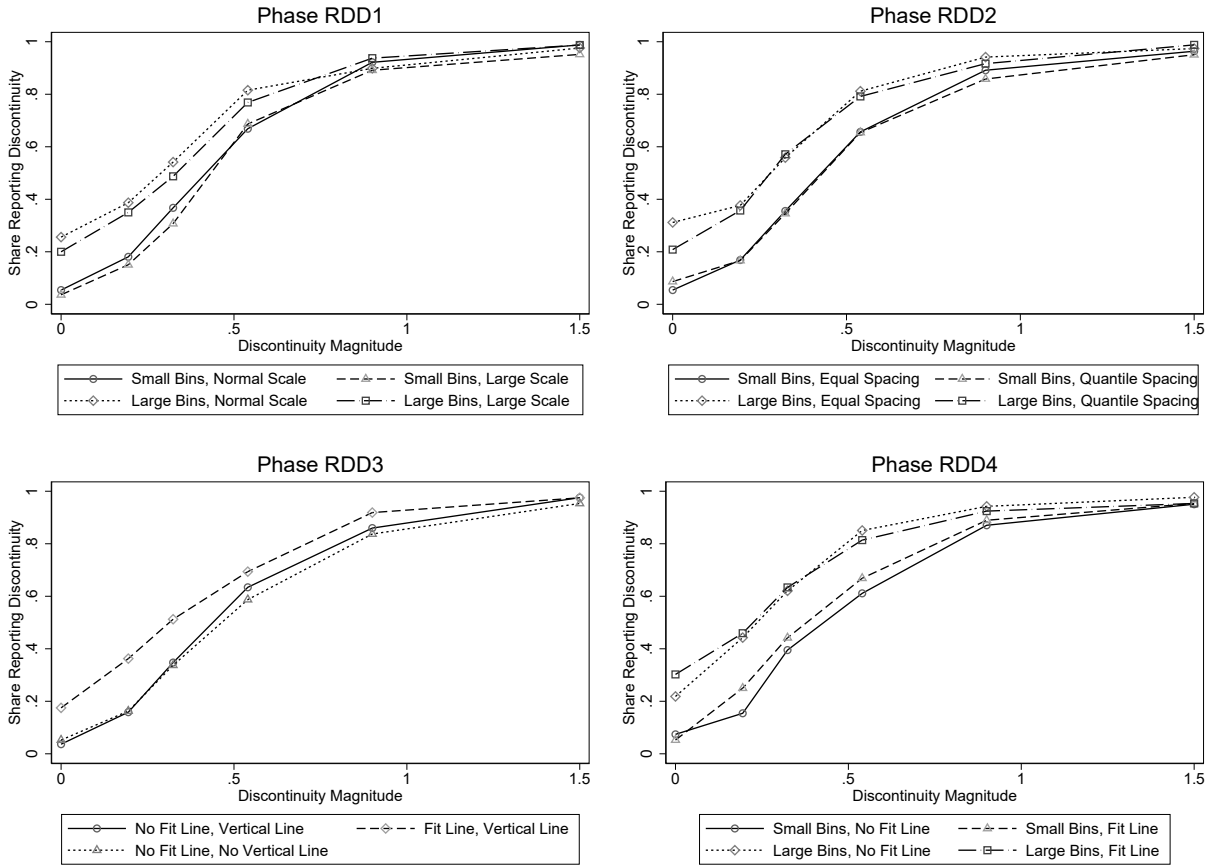
STOCK, J. H. AND M. W. WATSON (2011): *Introduction to Econometrics, 3rd edition*, Pearson.

WATSON, G. S. (1964): "Smooth Regression Analysis," *Sankhyā: The Indian Journal of Statistics, Series A*.

WOLFE, J. M. AND T. S. HOROWITZ (2004): "What Attributes Guide the Deployment of Visual Attention and How Do They Do it?" *Nature reviews neuroscience*, 5, 495–501.

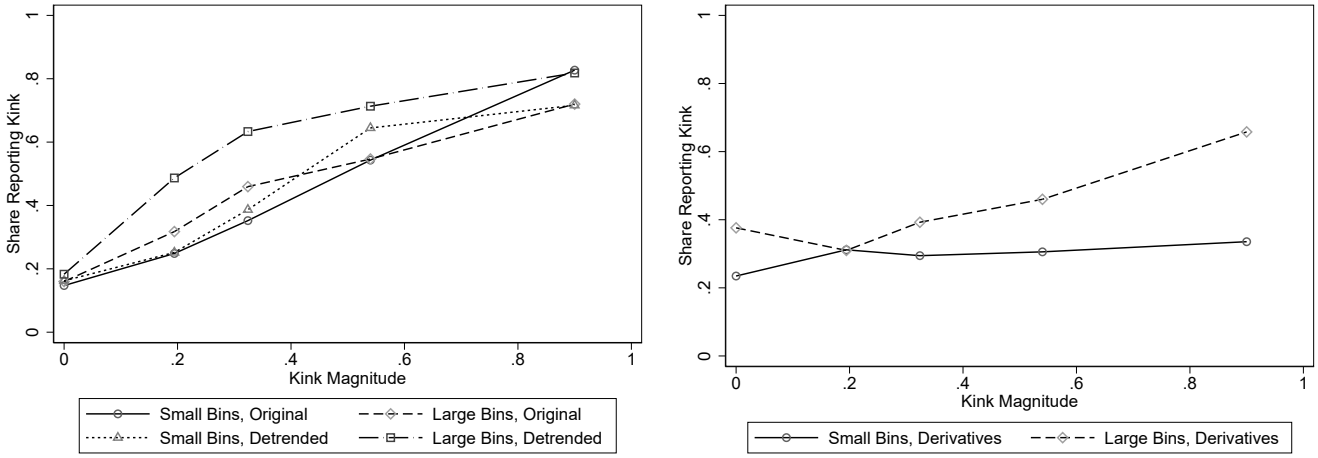
Figures

Figure 1: RDD Phases: Power Functions



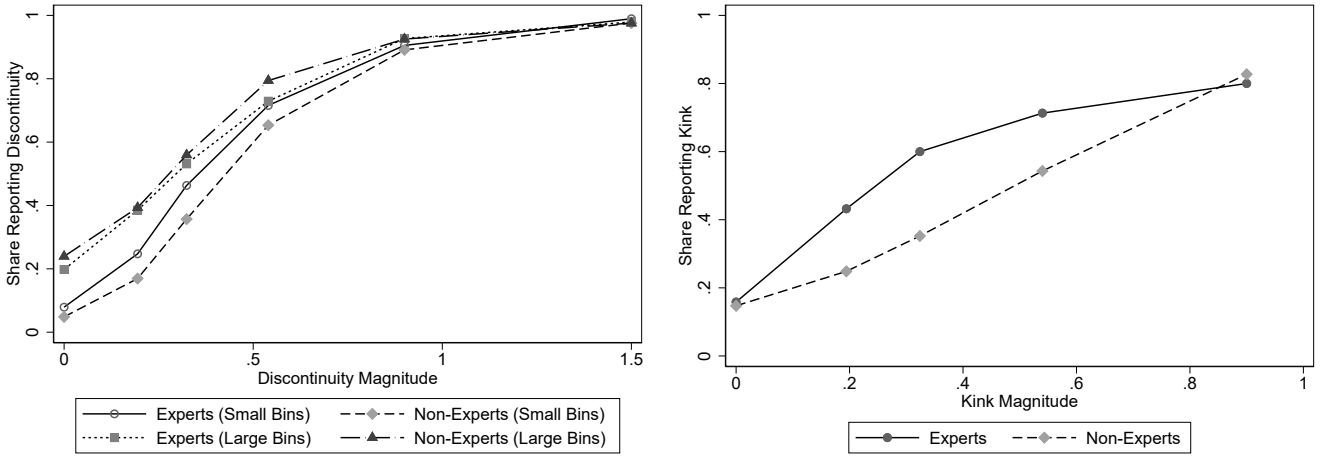
Notes: Plotted are empirical power functions from the four non-expert RDD experiments. The power functions are defined in Section 2. The discontinuity magnitude on the *x*-axis is specified as a multiple of the error standard deviation. The *y*-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure 2: RKD Phase: Power Functions



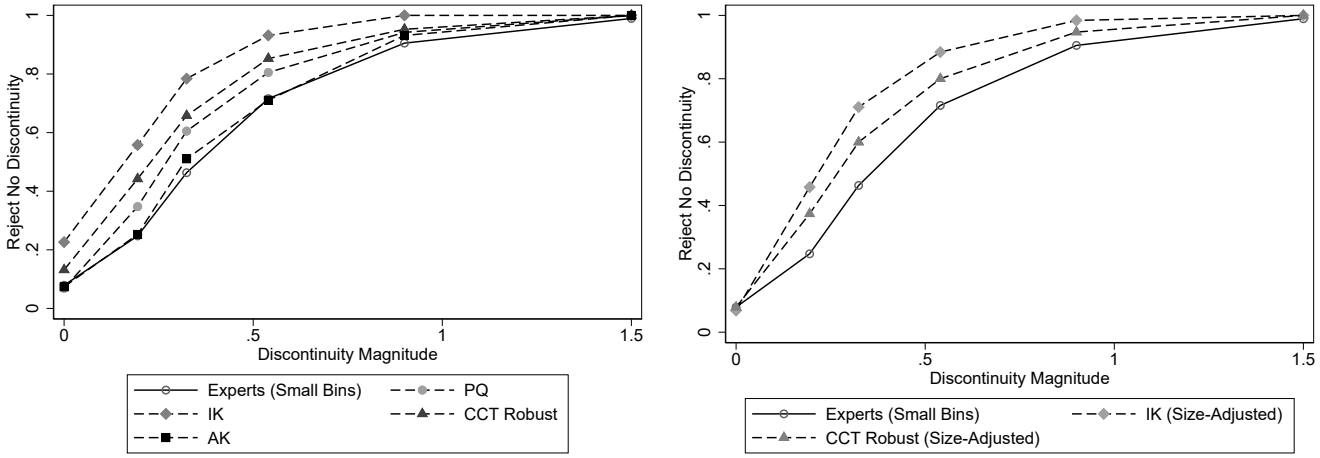
Notes: Plotted are empirical power functions from the RKD experiment. The power functions are defined in Section 2. The discontinuity magnitude on the x-axis is specified as the asymptotic t-statistics of the angle estimator between the left and right arms of the CEFs (see Section 3.2 and Appendix B for details). The y-axis represents the share of respondents classifying a graph as having a kink at the policy threshold. The left figure compares the power functions for the four RK treatments that plot the outcome in levels. The right figure compares the power functions for the two RK treatments that plot the first derivatives of the outcome with respect to the running variable.

Figure 3: Expert vs Non-Expert Performance



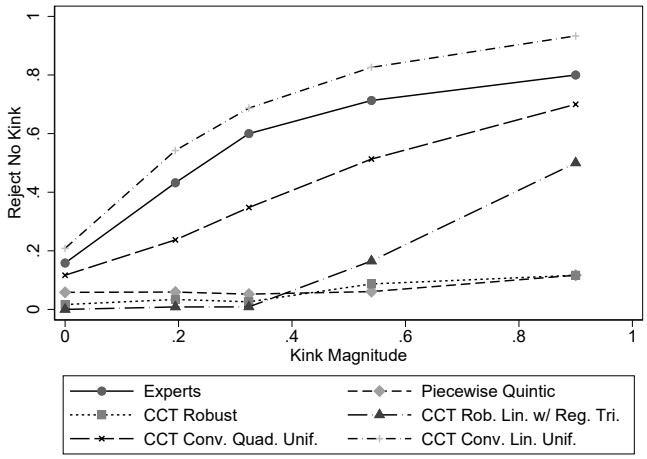
Notes: Plotted are the power functions for the experts and non-experts. The left and right figures compare their performances from reading RD and RK graphs, respectively.

Figure 4: RDD: Expert Visual vs Econometric Inference



Notes: PQ uses a correctly-specified regression model using piecewise quintics above and below the treatment threshold and assuming homoskedasticity. IK is based on a local linear estimator using the IK bandwidth. CCT Robust is the default RDD inference procedure from CCT's `rdrobust`, which uses the robust confidence interval based on a bias-corrected local quadratic estimator with triangular kernel and regularized (smaller) bandwidth. AK uses the `RDHonest` procedure with the true bound on each DGP's second derivative.

Figure 5: RKD: Expert Visual vs Econometric Inference



Notes: Piecewise Quintic is a correctly specified regression model using piecewise quintics above and below the treatment threshold and assuming homoskedasticity. CCT Robust is the default RKD inference procedure from CCT's `rdrobust`, which uses the robust confidence interval based on a bias-corrected local quadratic estimator with triangular kernel and regularized (smaller) bandwidth. CCT Rob. Lin. w/ Reg. Tri. is the same as CCT Robust, except it is based on a linear estimator. CCT Conv. Quad. Unif. is the inference procedure using the conventional confidence interval based on a uncorrected local quadratic estimator with uniform kernel and unregularized (larger) bandwidth. CCT Conv. Lin. Unif. is the same as CCT Conv. Quad. Unif., except it is based on a linear estimator.

Appendix (For Online Publication Only)

A Theory: Estimator Properties, Bin Selectors, and MSE Decompositions

A.1 Properties of the Type I and Type II Error Probability Estimators

Following Section 2, Proposition 1 states the properties of the g-specific estimator $\hat{p}(\gamma, g, d)$.

Proposition 1. *Under Assumption 1 and for a given DGP g*

1. $E[\hat{p}(\gamma,g,d)] = p(\gamma,g,d)$
2. $\hat{p}(\gamma,g,d) \xrightarrow{\mathbb{P}} p(\gamma,g,d)$ *as* $M \to \infty$
3. $M \cdot \hat{p}(\gamma,g,d) \sim Binomial(M, p(\gamma,g,d))$
4. $\sqrt{M}(\hat{p}(\gamma,g,d) - p(\gamma,g,d)) \Rightarrow N\left(0, p(\gamma,g,d) \cdot (1 - p(\gamma,g,d))\right)$.

The proof of the proposition follows trivially from Assumption 1.

Parts 1 and 2 of the proposition state that $\hat{p}(\gamma,g,d)$ is an unbiased and consistent estimator for $p(\gamma,g,d)$. Part 3 states that $\hat{p}(\gamma,g,d)$ has a scaled binomial finite sample distribution, for which methods such as the Clopper-Pearson confidence interval have been developed for finite sample inference, which allows us to construct confidence intervals on the g-specific type I and II error probabilities. The standard asymptotic normality result is provided in Part 4, but we do not invoke it in our analysis, as M is equal to eight in our experiments due to resource constraints.

Now we state the properties of the overall estimator $\hat{\bar{p}}(\gamma,d)$.

Proposition 2. *Under Assumptions 1 and 2*

1. $E[\hat{\bar{p}}(\gamma,d)] = \bar{p}(\gamma,d)$
2. $\hat{\bar{p}}(\gamma,d) \xrightarrow{\mathbb{P}} \bar{p}(\gamma,d)$ *as* $J \to \infty$
3. *Conditional on* $\{g_j\}_{j=1}^{J}$, $M \cdot J \cdot \hat{\bar{p}}(\gamma,d)$ *follows a Poisson binomial distribution*
4. *As* $J \to \infty$, $\sqrt{J}(\hat{\bar{p}}(\gamma,d) - \bar{p}(\gamma,d)) \Rightarrow N\left(0, var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d))\right)$, *with* $var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d)) = \frac{1}{M}\bar{p}(\gamma,d) + (1 - \frac{1}{M})E[p(\gamma,g,d)^2] - \bar{p}(\gamma,d)^2$.

Proof. Part 1:

$$\begin{aligned} E[\hat{\bar{p}}(\gamma,d)] &= \frac{1}{J}\sum_{j=1}^{J} E_{g_j\in\mathscr{G}}\left[E[\hat{p}(\gamma,g_j,d)|g_j]\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} E_{g_j\in\mathscr{G}}\left[p(\gamma,g_j,d)\right] \\ &= \bar{p}(\gamma,d) \end{aligned}$$

Part 2: Assumptions 1 and 2 imply that, for any J, $\{\hat{p}(\gamma,g_j,d)\}_{1\leq j\leq J}$ is a set of independent and identically distributed random variables. Because $\hat{p}(\gamma,g_j,d)^2$ is uniformly bounded between zero and one, the weak law of large numbers for triangular arrays (see, for example, Durrett, 2010) applies to the set $\{\hat{p}(\gamma,g_j,d)\}_{1\leq j\leq J, J=1,2,\ldots}$. Because the expectation of $\hat{p}(\gamma,g_j,d)$ is $\bar{p}(\gamma,d)$ for each j (Part 1 of the proposition), we have the desired result.[1]

Part 3: the statement simply follows the definition of a Poisson binomial distribution.

Part 4: Similar to the proof of part 2 of the Proposition, the result follows from an application of the Central Limit Theorem for Triangular Arrays (or the Lindeberg-Feller Central Limit Theorem):

$$\sqrt{J}(\hat{\bar{p}}(\gamma,d)-\bar{p}(\gamma,d)) \Rightarrow N\left(0, var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d))\right).$$

The law of total variance implies that

$$\begin{aligned} & var_{g\in\mathscr{G}}(\hat{p}(\gamma,g,d)) \\ =& E_{g\in\mathscr{G}}[var(\hat{p}(\gamma,g,d)|g)] + var_{g\in\mathscr{G}}(E[\hat{p}(\gamma,g,d)|g]) \\ =& E_{g\in\mathscr{G}}\left[\frac{p(\gamma,g,d)\left(1-p(\gamma,g,d)\right)}{M}\right] \\ & + var_{g\in\mathscr{G}}(p(\gamma,g,d)) \end{aligned}$$

and the last equality follows from Proposition 1. Because

$$var_{g\in\mathscr{G}}(p(\gamma,g,d)) = E_{g\in\mathscr{G}}[p(\gamma,g,d)^2] - E_{g\in\mathscr{G}}[p(\gamma,g,d)]^2$$

we have

$$var_{g\in\mathscr{G}}(p(\gamma,g,d)) = \frac{1}{M}\bar{p}(\gamma,d) + (1-\frac{1}{M})E_{g\in\mathscr{G}}[p(\gamma,g,d)^2] - \bar{p}(\gamma,d)^2$$

following simple algebra. □

[1] The application of the law of large numbers for triangular arrays, as opposed to its standard counterpart, allows $\hat{p}(\gamma,g_j,d)$ to change as J increases. Take $\hat{p}(\gamma,g_1,d)$, for example. We accommodate the scenarios where M, the number of participants who see g_1, changes with J, or where we allocate different individuals to see g_1 as we increase the number of DGPs sampled.

Parts 1 and 2 of Proposition 2 establish unbiasedness and consistency of $\hat{\bar{p}}(\gamma,d)$ as an estimator for $\bar{p}(\gamma,d)$. Part 3 provides the finite-sample distribution of $\hat{\bar{p}}(\gamma,d)$; the Poisson binomial distribution is that of a sum of independent but potentially non-identically distributed Bernoulli random variables. Part 4 states the asymptotic normality result for $\hat{\bar{p}}(\gamma,d)$ as $J \to \infty$. In principle, we can consistently estimate the variance and use Part 4 to conduct inference on $\bar{p}(\gamma,d)$ if J is large. As with M, however, we choose a moderate J (11) in our experiments due to resource constraints, and the asymptotic normality statement is, therefore, more conceptual than practical. Nevertheless, we make the following observations on the variance expression in Part 4 to help understand the variation in the estimator $\hat{\bar{p}}(\gamma,d)$. In essence, the variance expression reflects the block assignment nature of the DGPs to the participants. When $M = 1$, each DGP is only assigned to a single participant, the R_i's are i.i.d., and $\hat{\bar{p}}(\gamma,d)$ simply follows a (scaled) binomial distribution with variance $\bar{p}(\gamma,d) - \bar{p}(\gamma,d)^2$. When $M > 1$, each DGP is assigned to a block of participants, the R_i's are no longer independent within the same block, and the distribution of $\hat{\bar{p}}(\gamma,d)$ generally deviates from a (scaled) binomial. When M is large, we can precisely estimate $p(\gamma,g,d)$ for each DGP g, and the variance of $\hat{\bar{p}}(\gamma,d)$ reduces to that of $p(\gamma,g,d)$ over the space of DGPs.

In the actual analysis of our experimental data, we use the normal approximation of the binomial random variable $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ when $M \cdot J$ is large. Based on a result from Hoeffding (1956) and summarized by Percus and Percus (1985), using $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ leads to conservative inference on the Poisson binomial random variable $M \cdot J \cdot \hat{\bar{p}}(\gamma,d)$ when conditioning on $\{g_j\}_{j=1}^{J}$.[2] Therefore, conditioning on the set of selected DGPs, using the normal distribution $N\left(\hat{\bar{p}}(\gamma,d), \hat{\bar{p}}(\gamma,d)(1-\hat{\bar{p}}(\gamma,d))/(M \cdot J)\right)$ provides (approximately) conservative inference for $\hat{\bar{p}}(\gamma,d)$ and only relies on the product $M \cdot J$ being large, as opposed to J being large by itself.

A.2 Calonico et al. (2015) Bin Selection Algorithms

This section describes the two bin width algorithms from Calonico, Cattaneo, and Titiunik (2015): the IMSE algorithm, which minimizes the integrated mean squared error of the bin-average estimators of the CEF and results in fewer, larger bins, and the MV algorithm, which aims to approximate the variability of the underlying data and results in more, smaller bins.

Formally, consider without loss of generality the number of bins above the threshold. We denote the relevant quantities with a "+" subscript. The IMSE bin selector chooses the number of bins $J_{+,n}$ to minimize

[2] For a given (γ,d), inference for $\hat{\bar{p}}(\gamma,d)$ using $\text{Binomial}(M \cdot J, \bar{p}(\gamma,d))$ is exact when $p(\gamma,g_j,d)$ is the same for all j.

the integrated mean squared error

$$\int_0^{\bar{x}} E[(\hat{\mu}_+(x) - \mu_+(x))^2 | X = x] f(x) dx = \frac{J_{+,n}}{n} Var_+ \{1 + o_p(1)\} + \frac{1}{J_{+,n}^2} Bias_+ \{1 + o_p(1)\}$$

where μ_+ denotes the CEF, $\hat{\mu}_+$ denotes the bin-average estimator of μ_+, $f(x)$ is the density of X, n is the overall sample size, and the constants are equal to

$$Var_+ = \frac{1}{\bar{x}} \int_0^{\bar{x}} \sigma_+^2(x) dx$$

$$Bias_+ = \frac{\bar{x}^2}{12} \int_0^{\bar{x}} (\mu_+'(x))^2 f(x) dx,$$

with $\bar{x}$ being the upper bound of the support of X, μ_+' the first derivative of μ_+, and $\sigma_+^2(x)$ the conditional variance of Y given $X = x$.

Solving the minimization problem, we obtain the IMSE-optimal number of bins

$$J_{+,n,IMSE} = \left\lceil \left(\frac{2Bias_+}{Var_+} \right)^{\frac{1}{3}} n^{\frac{1}{3}} \right\rceil$$

with an analogous $J_{-,n,IMSE}$ for the number of bins below the threshold.

The goal of the mimicking variance bin selector, on the other hand, is to "choose the number of bins so that the binned sample means have an asymptotic (integrated) variability approximately equal to the amount of variability of the raw data" (Calonico et al., 2015). Consider again the problem of choosing the number of bins above the threshold. The MV criterion translates to setting

$$\frac{J_{+,n}}{n} Var_+ = Var(Y | X \geqslant 0) \equiv V_+.$$

It follows that the number of bins should be

$$J_{+,n} = \frac{V_+}{Var_+} n.$$

However, this number of bins grows too quickly for the rate conditions in Calonico et al. (2015) that ensure the consistency of $\hat{\mu}_+(x)$. Instead, the authors modify the formula slightly and use the following bin number

$$J_{+,n,MV} = \left\lceil \frac{V_+}{Var_+} \frac{n}{\log(n)^2} \right\rceil.$$

A.3 Mean Squared Error Decompositions

For each meta-graph (γ, g, d), let $d_i(\gamma, g, d, \varepsilon_i)$ be the discontinuity estimate by participant i. The "human estimate" of the discontinuity for this meta-graph is then given by the average $\hat{d}(\gamma, g, d) = \frac{1}{M} \sum_{i=1}^M d_i(\gamma, g, d, \varepsilon_i)$. Meanwhile, let $d_{\hat{\theta}}(g, d, \varepsilon_i)$ and $\hat{d}_{\hat{\theta}}(g, d) = \frac{1}{M} \sum_{i=1}^M d_{\hat{\theta}}(g, d, \varepsilon_i)$ be the corresponding estimates given by the

econometric estimator $\hat{\theta}$. In order to better compare the performance of the humans versus the econometric estimator, we further decompose the mean squared error (MSE) of the above estimates into the square of the bias and the variance, by using the following identities respectively:

$$\frac{1}{M}\sum_{i=1}^{M}(d_i(\gamma,g,d,\varepsilon_i)-d)^2=(\hat{d}(\gamma,g,d)-d)^2+\frac{1}{M}\sum_{i=1}^{M}(d_i(\gamma,g,d,\varepsilon_i)-\hat{d}(\gamma,g,d))^2 \tag{A1}$$

$$\frac{1}{M}\sum_{i=1}^{M}(d_{\hat{\theta}}(g,d,\varepsilon_i)-d)^2=(\hat{d}_{\hat{\theta}}(g,d)-d)^2+\frac{1}{M}\sum_{i=1}^{M}(d_{\hat{\theta}}(g,d,\varepsilon_i)-\hat{d}_{\hat{\theta}}(g,d))^2. \tag{A2}$$

Note that the interpretation of the quantities in the first identity differs slightly from the conventional MSE decomposition, as the randomness here not only comes from the realization (ε_i), but also from the participant (d_i). Nevertheless, the second term can still be conveniently understood as the variance, as $d_i(\gamma,g,d,\varepsilon_i)$ remains independent and identically distributed for different i. In Figure A.20, we plot the square root of the expression (A1) and the expressions (A2) for various $\hat{\theta}$. We also decompose the difference between the left-hand sides of (A1) and (A2) into differences in bias and variance for the case where $\hat{\theta}$ is the piecewise quintic estimator.

B Additional Details on DGP Specification and Graph Creation

In conjunction with Section 3.2 in the main text, this appendix describes the process by which we specify DGPs and create graphs for our experiments.

For RDD, we identify 110 empirical papers drawn from top economics journals including the *American Economic Review*, *American Economic Journals*, *Econometrica*, *Journal of Business and Economic Statistics*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics and Statistics.* We then randomly sample 11 papers from this list that have replication data available.

For the specification of the running variable for each DGP, we directly use the empirical distribution of X from the corresponding paper and normalize the running variable to lie in $[-1,1]$. If the support of the running variable is asymmetric, this normalization may create a discontinuity in the density at $x=0$, which violates the assumptions in the identification framework of Lee (2008) (but not the minimally sufficient identifying assumptions of Hahn et al., 1999). The original running variable in seven of our 11 papers fails the McCrary (2008) test for continuity of the running variable density. Among our normalized running variables, eight fail the test, with two passing only in their non-normalized version and one passing only

after normalizing. In our testing, visual inference performance does not vary significantly across DGPs whether the support of their original (unnormalized) running variables is symmetric. Following Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014), we remove any observations where $|X| > 0.99$.

Next, we use the resulting datasets and further follow Imbens and Kalyanaraman (2012) and Calonico et al. (2014) to specify the DGP's CEF via a piecewise global quintic regression. Denoting the left and right intercepts of the CEFs by α^- and α^+, we shift the right arm of the CEF upwards/downwards by the amount $|\alpha^+ - \alpha^-|$ to make the functions $E[\tilde{Y}|X = x]$ continuous at the policy threshold.

Our specification of the distribution of the error term u as i.i.d. normal similarly follows Imbens and Kalyanaraman (2012) and Calonico et al. (2014). u has mean zero and standard deviation σ, which is specified as the root mean squared error of the global quintic regression in the CEF specification above. We generate eight draws of u for each DGP, enough so that every graph we generate for our experiments is seen by no more than one participant. Most of our RD papers use a continuous running variable, but two feature semi-discrete variables where the running variable takes on many unique values but these points have multiple observations each. In these two cases, we add small amounts of noise from a $N(0, (\frac{1}{\min\{N_-, N_+\}})^2)$ distribution, where N_- and N_+ are the number of observations below and above the policy threshold, to match the continuous running variable condition assumed in Calonico et al. (2015). We add this noise prior to fitting the piecewise quintic CEF. Perturbing the data prior to fitting the quintics introduces "measurement error" that can attenuate estimated discontinuities. Alternatives include fitting the CEF prior to adding the noise or drawing the noise from a uniform distribution to prevent observations from crossing the policy threshold. For a thorough discussion of measurement error in RD, see Pei and Shen (2017).

In practice, the difference between adding the noise before and after fitting the CEF is minor. Figure A.22 presents lineup protocols for these two DGPs. To test the null hypothesis that the DGP resulting from fitting the piecewise quintic prior to adding noise is indistinguishable from the DGP where noise is added first, we randomly place one graph from the former distribution among 19 from the latter. All graphs use a continuous running variable generated by adding noise to the original semi-discrete one. The solutions, i.e. the graphs from the DGP fit on the original data, are in the footnote at the end of this sentence.[3] We present both sets of CEFs in Figure A.23, with the scale of the y-axis chosen to match that of the corresponding

[3]The solution for the left lineup is row $5 \cdot 2 - 6$ and column $2^2 - 1$. The solution for the right lineup is row $-3 \cdot 2 + 10$ and column $9/3 - 2$.

graph in the original paper. Although the differences in the tails for the CEF on the right of Figure A.23 underscore the sensitivity of the piecewise quintic specifications, the difficulty of these lineup protocols and the general similarity of the CEFs suggest that the infusion of measurement error is not a major concern.

Specifying the RKD DGPs requires several modifications to our process. As mentioned in Section 3.2, we generate the running variable by sampling from an estimated kernel density due to data confidentiality issues. We again normalize the support of the running variable to $[-1,1]$ with observations where $|X| > 0.99$ trimmed. Because the X variable in all of the RKD applications is income, its distribution is skewed and its support quite asymmetric around 0. Therefore, we select the largest subsample such that the support of X therein is symmetric around 0.

While one can create arbitrary discontinuities that preserve the shape of the CEF just by vertically shifting one of the DGP's "arms," kinks are not as simple. How do we create a zero-kink graph from the seven kinked CEFs in Figure A.7? (The four global quintic CEFs already have no kink.) We decide to rotate the right arm of the CEF until the left and right derivatives match. Following that logic, we create graphs with kinks by rotating the right arm from the zero-kink graph at nine different levels. As explained in Section 3.2, we create the four smooth global quintic DGPs out of concern that only using the seven piecewise quintic DGPs may be too favorable to visual inference. Specifically, several of the zero-kink CEFs of the piecewise quintic DGPs obtained by rotating their right arms appear similar to a linear function, at least locally around $x = 0$. In comparison, the curvature in the global quintic DGPs, especially in combination with the function's peaking at $x = 0$, is likely much harder to distinguish from a kink for visual inference.

Because we generate the RK graphs by rotations, we express the kink magnitude in terms of the angle between the left and right arms of the CEF. A kink is defined as the difference between the right derivative (denoted by κ^+) and the left derivative (denoted by κ^-), $\kappa^+ - \kappa^-$, and the corresponding angle is simply $\psi \equiv \arctan(\kappa^+) - \arctan(\kappa^-)$. Our nine rotations are $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$ radians, where ξ is the asymptotic standard deviation of the angle estimator between the left and right arms $\hat{\psi} \equiv \arctan(\hat{\kappa}^+) - \arctan(\hat{\kappa}^-)$ and $\hat{\kappa}^+$ and $\hat{\kappa}^-$ are estimators of κ^+ and κ^- from a piecewise quintic regression on the data used to calibrate the CEF. A complication of this rotation-based approach is that the right arm of a function can be rotated only a certain amount before it bends back over itself and ceases to be a function. For each RK DGP, we calculate the maximum angle such that its rotated CEF remains a function, and this maximum prevents us from specifying the largest kink magnitude as $\pm 1.5\xi$ for some of the DGPs. While the maximum kink magnitude in our RK DGPs is $\pm 0.9\xi$, we cannot create kinks that large for all of

the DGPs. We replace two meta-graphs that would go too far, $+0.54\xi$ and $+0.9\xi$ for Card, Lee, and Pei (2009), with their non-offending negative counterpart rotation.

Specifying the distribution of the error term u is more complex in RKD for two reasons. First, adding u prior to rotation may result in points changing order along the x-dimension post-rotation, exacerbating the maximum rotation problem. Second, unlike the vertical shift technique used in creating RD datasets and graphs of different discontinuity levels, data rotation changes the standard deviation of the error term. To circumvent these problems, we rotate the right arm of the CEF first and then add the error term to generate the outcome variable. The standard deviation of u depends on the degree of rotation and is calculated such that ξ, the asymptotic standard deviation of the angle estimator $\hat{\psi}$ in the original fitted CEF, remains the same after rotation. Like in the RD case, we use eight draws of u for each DGP.

C Design of Experiments and Studies

This section outlines the structure of the RDD and RKD experiments. We first describe the general sequence of events and survey structure before discussing the sequential rollouts of the RDD and RKD experiments. All parts of this study were approved by the Institutional Review Boards (IRBs) of Cornell University, Princeton, Columbia Teachers College, and the University of Waterloo and were pre-registered on Open Science Framework and at the American Economic Association's RCT Registry. Expert studies run as part of seminars were additionally approved by their respective local IRBs. Before beginning any of the studies, participants completed consent forms. Participants had the opportunity to exit the study at any time.

C.1 Non-Expert Experiment

Each experiment consisted of three parts. In Part 1, participants watched a video tutorial on RDD (RKD) outlining the graph construction (normalization of the running variable, binning the data, etc.) and stating the objective as classifying a discontinuity (kink) in the "true" underlying relationship at $x = 0$. The video tutorial lasted between three to four minutes depending on the participant's graphical representation treatment. For example, participants in treatment arms featuring fit lines received a brief supplementary segment informing them that fit lines only serve as approximations of the true underlying relationship between the running variable and the outcome variable. To ensure participant engagement, the video tutorial featured an attention check about halfway through the video. This took the form of a colored bird (see example

in Figure A.24); the voiceover on this slide informed participants that when asked about the color of our "bird of interest," they should report a specific color that did not match the color of the bird in the picture. For example, participants seeing Figure A.24 below might be asked to report that the bird is red, not blue. Participants had the opportunity to watch the video several times or to pause and rewatch specific sections as desired. However, participants were not able to return to the video tutorial once they had moved on to the next section.

The video tutorial was followed by a sequence of example questions. Participants were asked to classify tutorial graphs and received feedback on their answers. Participants could navigate between these graphs using a panel of buttons on the right-hand side of the screen (see Figures A.25 and A.26).

Part 2 of the experiment consisted of a series of paid classification tasks. Participants were asked to assess whether a given graph featured a discontinuity/kink or not and could choose between two potential bonus options for each graph. We discuss this bonus system in detail in the next subsection. Figure A.27 shows a typical classification screen. Participants were not given any feedback regarding their accuracy between tasks.

Part 3 of the experiment was an exit survey. We solicited basic sample demographics such as age, gender, occupation, education level, and prior experience with statistics. For all phases after the initial pilot, we also asked participants to report whether they had participated in earlier iterations of the study. At the end of the experiment, participants were informed of the total number of graphs they classified correctly as well as their corresponding bonus earnings (see Figure A.28 for an example). Experiment participants were paid in Amazon gift cards after each phase was complete.

C.1.1 Analysis of Bonus Payments

Participants in the experiment had the option to choose between earning a bonus of 40 cents if their discontinuity/kink classification was correct and nothing otherwise, or to earn a bonus of 20 cents irrespective of their classification. Note that participants will choose a given answer if their perceived probability that the answer is correct is at least 50%. Denoting the utility function by u, the expected utility under the wager for a perceived probability of $p \equiv \mathbb{P}(correct)$ is given by $pu(\$0.40) + (1-p)u(\$0)$, which is equal to $pu(\$0.40)$ if we normalize $u(0)$ to zero. Assuming expected utility maximization, participants will choose the sure amount whenever $pu(\$0.40) \leq u(\$0.20)$, i.e. whenever $p \in [0.5, u(\$0.20)/u(\$0.40)]$. Further assuming that subjects are approximately risk neutral at small stakes, participants would only choose the sure

amount whenever they believe their chance of being correct to be a coin flip (because $u(\$0.20)/u(\$0.40)$ is approximately equal to 0.5 in this case).

If we instead allow for loss aversion, participants prefer the sure amount under a wider range of (perceived) success probabilities. Under loss aversion, the structure of the model is similar except that the utility function is replaced with a value function for gains and losses with respect to a reference point (Kahneman and Tversky, 1979). Prospect theory is based on the assumption that losses loom larger than gains, which is commonly operationalized through a loss-aversion parameter $\lambda > 1$ multiplying outcomes in the loss domain. Assuming that participants view the sure amount of 20 cents as their reference point, prospect theory suggests that they would consider the gamble as an opportunity to win 20 cents or lose 20 cents depending on their answer. The expected payoff of the wager in this case is given by $pu(\$0.40 - \$0.20) - \lambda(1-p)u(\$0.20 - \$0) = u(\$0.20)(p - \lambda(1-p))$, which would be compared to a certain outcome of $u(\$0.20 - \$0.20) = 0$ under the fixed payment option. Participants would therefore choose the wager whenever $p/(1-p) \geq \lambda$. Estimates for λ commonly range between 1.5 and 2.5, though significant heterogeneity may exist with respect to this preference parameter. See Chapman et al. (2018) for a review.

C.1.2 RDD

The rollout of the RDD experiments was staggered across four phases which were completed between November 2018 and October 2019. The first phase, phase RDD1, served as a pilot and was not part of the pre-registration package. The goal of this pilot was to gauge effect sizes in the case of the bin width and axis scaling treatments in order to determine the required sample size for subsequent (pre-registered) phases. Table A.2 outlines the graphical representation treatments implemented in each phase. Each participant's classification task featured 11 graphs, one from each of the 11 specified DGPs. The order of graphs was randomized within subjects. The discontinuity levels were chosen as follows: two graphs featured no discontinuity, one graph featured the extreme discontinuity levels of $\pm 1.5\sigma$, and the remaining eight graphs were based on each of the eight discontinuity levels $\pm 0.1944\sigma, \pm 0.324\sigma, \pm 0.54\sigma, \pm 0.9\sigma$. In total, each treatment arm contained 968 unique graphs split across 88 unique participants. Figure A.29 outlines the sequence of events for the RDD experiments.

C.1.3 RKD

The rollout of the RKD experiments consisted of a pilot followed by a single experiment, both completed between April and June of 2019. Table A.3 outlines the graphical representation treatments implemented in each phase. Each classification task was split into two distinct sections. One section corresponded to RKD graphs presented in levels and required participants to assess the presence of a kink. The other section instead plotted estimated derivatives and required participants to assess the presence of a discontinuity. The order of the two sections was randomized between participants.

In the RKD pilot, all graphs were based on 7 piecewise quintic DGPs. Participants saw one graph from each DGP in each of the two sections, for a total of 14 graphs. In the final experiment (Phase RKD), we added four additional global quintic DGPs that featured smooth peaks instead of kinks at $x = 0$ and asked participants to classify a total of 22 graphs (11 graphs in each section). The test kink levels (measured in radians) were $0, \pm 0.1944\xi, \pm 0.324\xi, \pm 0.54\xi, \pm 0.9\xi$, for a total of nine possible levels. The pilot featured 504 unique graphs for each treatment arm split across 72 unique surveys, and Phase RKD featured 792 unique graphs per treatment arm split across 72 unique surveys. Figure A.30 outlines the sequence of events for the RKD experiments.

C.2 Expert Study

The expert study was run both online using members of the National Bureau of Economic Research in applied microeconomic fields (aging, children, development, education, health, health care, industrial organization, labor, and public) and members and affiliates of the Institute of Labor Economics (IZA), and during three seminar presentations between May and October of 2019. Table A.11 shows the timeline and number of participants for each session. Figure A.31 outlines the sequence of events for the expert study.

Unlike in the non-expert experiments, participants in the expert study were not asked to watch a video tutorial or complete practice questions at the start. Instead, Part 1 of the expert study was a classification task paralleling the one non-expert participants completed (see Figure A.33). In addition, the graphical representation choices in this study were not randomized between subjects. Participants had the opportunity to navigate between the 11 graph classifications using a panel of navigation buttons at the bottom of the screen. This feature was introduced to address feedback we received during the testing of the survey. Figure A.33 shows an example of the task screens for participants in the RD and RK treatments.

Part 2 of the expert study asked participants to rank four alternative graphical representation options for RDD or two for RKD. The four treatment combinations in the RD case corresponded to small or large bins interacted with both fit line treatments, as in Phase RDD4. The two treatment combinations in the RK case were small and large bins. The instructions for Part 2 of the expert survey are outlined in Figures A.34 and A.35. Participants were asked to specify which graphing treatment they believe researchers should use to present evidence of the main treatment effect of a study, as well as which graphical options they believe would perform best and worst in our non-expert sample. Participants were also able to report having "no preference" across the different graphing options. These three decisions were repeated three times in the context of a zero, small, and large discontinuity or kink.

Part 3 of the expert study was an exit survey. We asked about basic sample demographics such as age, gender, geographic region, main area of research, and prior experience with RD/RK. After all sessions of the expert study were complete, we randomly selected four participants to receive a participation fee of $450 plus $50 for every correctly classified graph in Part 1 of the study. We did not take the accuracy of the discontinuity magnitude estimate into account to determine payments in the RD case, only the binary classification decision.

Table A.11 highlights the different participant pools for the sessions of the expert study and lists the graphical representation choices in the classification task (Part 1 of the expert study). All graphs in Part 1 used the default Stata axis scaling with no fit lines and a vertical line at $x = 0$. All but one session of the expert study used small bins for the RD case. The exception was the session run at Princeton's Quantitative Social Science Colloquium in October 2019 in which we used large bins to compare the effect of bin widths on experts and non-experts.

The distribution of graphs seen by the experts and non-experts is slightly different due to the methods used to randomize participants' graphs. The non-expert experiment's randomization is based around the order in which the 88 participants for each treatment arm begin the survey. We use the same randomization mechanism for experts, but repeating the survey in different seminars and online phases results in more experts at the "beginning" of the graph assignments. Reweighting the data to match the distribution of graphs seen by non-experts does not meaningfully change any of our results.

C.3 Eyetracking Study

We chose the 11 graphs for the eyetracking study participants to classify based on information available when designing it. Specifically, we had data from the first three phases of experiments on RD graphical parameters as well as expert responses from two seminars. We saw the experts in these early samples attain both lower type I and type II error rates than the non-experts, and we chose the discontinuity-DGP pairs based on the maximum difference in classification accuracy between experts and non-experts. The final set of 11 graphs features four graphs with no discontinuity, three graphs with a discontinuity magnitude of 0.1944σ, two graphs with a discontinuity magnitude of 0.324σ, and two graphs with a discontinuity magnitude of 0.54σ. Like those shown to most of the participants in the expert study, all graphs have equally spaced small bins and default Stata axis scaling with a vertical line indicating the policy threshold and no fit lines.

The eyetracking study consisted of three parts for non-experts and two for experts (see Figure A.36 for details). For non-experts, Part 1 was the same video- and example-based tutorial used in the RDD experiments. To avoid eye fatigue, this tutorial was completed on a regular computer. After completing the tutorial, participants were moved to a Tobii XL60 eyetracker to complete Part 2, which consisted of 11 graph classification tasks. Each task was broken up into three separate screens as illustrated in Figure A.37: a white screen with a cross around the perimeter, a screen showing the graph of interest, and a screen asking participants to report whether or not they believe the graph featured a discontinuity. Presenting a white screen with a cross avoids "central fixation bias": by diverting participants' gaze towards the perimeter of the screen before showing a graph, we can rule out that time spent looking at the cutoff is simply an artifact of people's tendency to fixate their gaze at the center of the screen when a new stimulus appears. When looking at the graph of interest, participants could spend as much time as desired before proceeding to the next screen to answer the classification question. Participants could advance between pages by pressing the space bar in order to avoid having participants look down at the keyboard or having them use a button on the screen, which might confound their viewing patterns. Part 3 consisted of verbal feedback on the number of correct classifications and the corresponding bonus payment. Experts moved directly to the classification task at the eyetracking terminal without having to complete the tutorial.

To minimize distractions, participants completed the study one-by-one. Participants were paid in Amazon gift cards after each session.

D Subject Characteristics and Performance Predictors

D.1 Demographic Balance and Effects

For our non-expert experiment, we check the validity of our randomization by testing for the balance of covariates across treatment groups. For each phase, we regress the covariates on treatment group indicators and test whether all coefficients are equal. Specifically, we test for the balance of sex, education, age categories, statistical knowledge, passing the attention check, being a first-time participant, and not completing the experiment. The results are in Tables A.12 through A.15. Across these 44 hypotheses, we cannot reject balance in all but two instances (4.5% of tests) when using a 5%-level test: college completion and fraction of participants aged 23-49 in Phase RDD4. In addition, p-values for joint tests of significance for all covariates within each phase based on Pei, Pischke, and Schwandt (2019) are 0.18, 0.50, 0.65, and 0.14 for Phases RDD1-4 and 0.54 for Phase RKD. The evidence is consistent with successful randomization.

As can be seen in Tables A.17 and A.18, participant demographics such as gender, age, college completion, statistical knowledge, and being a repeat participant are generally not predictive of performance in the experiment for both discontinuities and kinks, with insignificant point estimates of a few percentage points. Out of the 34 factors we test, two have a statistically significant effect on the probability of classifying a graph correctly (5.9% of tests). In Phase RDD1, failing the attention check is associated with a 5.5pp decrease in the probability of classifying a graph correctly. This effect is smaller and insignificant in Phases RDD2 (4.3pp), RDD3 (-1.2pp), and RDD4 (-1.2pp). In Phase RDD3, being 50 or older correlates with a 9.4pp increase in the probability of classifying a graph correctly. This lack of predictive strength makes any imbalances in demographics across phases an even smaller concern. It also suggests that comparisons of treatments across phases are unlikely to be confounded by differing subject pools.

D.2 RD DGP Characteristics

While our results in Section 4 focus on the treatment effects of each graphical parameter, in this section we focus on other factors of inference performance. In particular, we explore how the characteristics of an RD DGP affect visual inference.

Figure A.38 plots the density of each RD DGP along its running variable. These distributions are varied. Some are approximately uniform, some are more normal, and others, such as DGPs 6 and 8, are highly skewed. Some DGPs show clear changes in density at the policy threshold, some of which were present in

the original data and some of which were introduced by having asymmetric supports prior to our rescaling the running variable. Figure A.39 compares power functions for the eight DGPs with symmetric supports prior to normalization with the three with asymmetric supports in Phase RDD4. There is a noticeable flatness or dip in the share of participants reporting a discontinuity between 0 and 0.1944σ. This pattern appears in every RD phase, but with only three asymmetric DGPs, this may also be due to the effects of the individual DGPs rather than something about the effects of normalizing the running variable when asymmetry is present.

When the distribution of the running variable is uniform, there is no difference between graphs generated with even spacing and quantile spacing. When the distribution is not uniform, we may expect quantile spacing to perform better, for example by preventing outlier bins with very few observations in more sparsely populated regions of the support. To quantify a DGP's deviation from a uniform distribution, we can compute Gini coefficients based on the distribution of observations within evenly spaced bins.[4] We calculate Gini coefficients for both the IMSE and MV bin width algorithms and take the average across ε for each RD DGP. Figure A.40 explores how the Gini coefficient interacts with bin spacing to test whether quantile spacing outperforms equal spacing when the distribution is more skewed by averaging the probability of a correct response across treatment arms within each DGP. There is no relationship between a DGP's Gini coefficient and performance across spacing types. Among DGPs with higher distributional inequality, as well as for those with intermediate and lower values, neither spacing dominates the other.

In Table A.19, we examine how the first derivatives of each arm of the DGP interact with the direction of the discontinuity in impacting power. Specifically, we ask whether visual inference performs better if the discontinuity is positive or negative within a combination of left and right slope signs. Looking only at graphs with nonzero discontinuities, with which we measure power, we regress the binary classification decision on dummy variables for the negative discontinuity sign within each slopes combination (omitting the positive discontinuity dummy) with DGP fixed effects and clustering at the participant level. These regressions show that visual inference performs better when the discontinuity direction matches the signs of both the left and right derivatives. For example, when the left and right first derivatives are both positive, power is 11 percentage points worse when the discontinuity is negative than positive (see Figure A.41 for an illustration). Table A.20 repeats a similar exercise for size over the no-discontinuity graphs. We cannot

[4]We do not compute Gini coefficients with quantile spacing, as all bins have about the same number of observations and the Gini coefficients would all be approximately zero.

include DGP fixed effects here, as they would eliminate the variation needed to identify impacts. Unlike with power, we do not find any meaningful results for size.

D.3 Dynamic Visual Inference

In this subsection, we investigate participants' performance over the course of the 11 graphs they see. This exercise provides evidence to alleviate the concern that participants have a preconceived notion that about half of the graphs feature a discontinuity, which may bias our results in favor of good size control by RD visual inference.

First, we trace out the fraction of participants reporting a discontinuity over their 11 graphs. If there is a consistent upward or downward trend, it would be indicative of participants having a fixed number of continuous graphs in mind. As shown in Figure A.42, however, the trends are flat in the treatment arm (small bins and no fit lines) used to compare visual and econometric inference procedures over the four RDD non-expert phases and in the expert sample.[5]

We also implement an AR(1) regression to more closely examine the dynamic discontinuity classification over the 11 graphs, which Figure A.42 may miss:

$$R_{is} = \rho R_{i(s-1)} + c_i + u_{is}. \tag{A3}$$

R_{is} is the binary variable indicating whether participant i reports a discontinuity in graph s, c_i is the individual fixed effect, and u_{is} is the population residual from the projection of R_{is} on $R_{i(s-1)}$ and c_i. We expect a negative ρ if participants believe ex-ante that half of the graphs will feature a discontinuity. In fact, we do see negative estimates of ρ ranging between -0.1 and -0.2 as reported in column (1) of Table A.21. However, these estimates are subject to a large-N asymptotic bias of order $O(1/S)$ per Nickell (1981) where S is the number of graphs. Inverting equation (18) from Nickell (1981), we solve for the bias-corrected estimates of ρ and report them in column (2) of Table A.21. Although these estimates are still negative, they are considerably closer to zero. For the four phases of the non-expert Phases RDD1-4, three bias-corrected estimates are statistically insignificant at the 5% level, while the other one is marginally significant.

Although the estimate of ρ in the expert sample is still significant after bias correction, the fact that it is the largest in magnitude of the five estimates in Table A.21 is reassuring. Because non-experts appear to

[5]Because we randomize the graph order, the flat discontinuity classification rate is consistent with a flat correct classification rate, which we also verify in the data. That is, no evidence supports "learning" over the 11 graphs in terms of visual inference success, which is concordant with our design in which participants find out the total number of correct discontinuity classifications only after they complete the study.

exhibit better size control than experts and are less experienced with RDD graphs, we may be more concerned that the non-experts' performance is driven by having a fixed number of continuous graphs in mind. However, this is refuted by the four estimates of ρ from the non-expert sample being smaller. Therefore, we conclude that the evidence does not support the hypothesis that participants expect half of the graphs to be continuous, and the good size control of RD visual inference is not a result of it.

E What *t*-Statistic Can Eyes Detect?

Using our experimental results, we can provide an evidence-based answer to the question of how large a t-statistic needs to be in order for readers to visually detect a discontinuity (a question that was also posed by Kirabo Jackson on Twitter). We do so by using an alternative scaling of the discontinuity. Instead of specifying the discontinuity d as a multiple of the error standard deviation σ as in Section 4, we specify the x-axis of the power functions in Figures A.43 and A.44 as $d/\sigma\sqrt{(X'X)^{-1}_{dd}}$, where X is a regressor matrix containing the 12 polynomial terms of the running variable for a piecewise quintic regression and the dd subscript denotes the entry of the matrix corresponding to the discontinuity estimator. Because the denominator is the large-sample error of the discontinuity estimator from a (correctly specified) piecewise quintic regression, the x-axes of these figures have a t-statistic interpretation. Note that because the rescaling results in a distinct set of six discontinuity magnitudes for each of our 11 DGPs, we now have 66 discontinuity magnitudes. Therefore, we bin points along the x-axis, and consequently, these power functions are no longer monotone, as the composition of DGPs changes across bin points.

Our analysis shows that the power of visual inference approaches 0.8, i.e. 80% of the participants correctly report a discontinuity, across all phases of our RD experiment as the t-statistic from the correctly specified regression exceeds 4. There are caveats to this already nuanced answer. First, this threshold t-statistic is specific to the range of parameters we test. For example, one could substantially increase the y-axis scale, which would be sure to make any discontinuity we test in this paper disappear before the human eye. Second, as we see in the RK results, this threshold is specific for detecting discontinuities, and a much smaller threshold applies to kink detection.

One might conjecture that the t-statistic threshold is sensitive to the sample size. In particular, it seems possible to keep the plot fixed and increase the sample size arbitrarily by adding observations in each bin, which would increase the t-statistic without altering visual inference. The pitfall of this argument is that for

the plot to look the same while increasing sample size, one needs to change σ as well. In fact, σ needs to increase proportionally to $\sqrt{n}$, leading to an invariant t-statistic.

F Combining Visual and Econometric Inferences

In section 4.4, we present "marginal" power functions for visual and econometric inference procedures. We now discuss the joint distribution of visual and econometric discontinuity/kink tests. Knowing the joint distribution i) informs us of their complementarity, i.e. whether visual and econometric inferences tend to be wrong on different data, and ii) is useful for combining inferences.

We begin by considering empirical results on i). From the data that underlie the power function comparisons in Figures 4 and 5, we produce a series of two-by-two contingency tables for the binary visual inference decision R and the binary econometric inference decision based on an estimator $\hat{\theta}$, $R_{\hat{\theta}}$. More specifically, for each normalized discontinuity or kink magnitude $|d|$ ($|d|$ takes on the values 0, 0.1944, 0.324, etc. per the design of our experiment), we characterize the joint distribution of R and $R_{\hat{\theta}}$ over the set of graphs presented to the expert participants. We conduct Fisher's exact test for independence and present one-sided p-values in Table A.22 by estimator $\hat{\theta}$ and discontinuity/kink magnitude $|d|$.[6]

Several patterns emerge from Table A.22. For RDD (Panel A), the first row shows no strong support for an association between econometric and visual inference decisions when the true discontinuity is zero. But when the true discontinuity is nonzero (rows two to six), there appears to be strong evidence in support of association (though not reported in Table A.22, all correlations are positive in these cases). In other words, type II errors by experts are predictive of type II errors by various econometric inference methods, but this is not true for type I errors. For RKD (Panel B), the evidence for an association is much weaker. However, consistent with our interpretation of Figure 5, the p-values generally become smaller and the evidence of positive association becomes stronger when we move from the default `rdrobust` inference (labeled CCT1) to conventional local linear inference with a relatively large bandwidth (labeled CCT4).

To addresses point ii), we provide a conceptual discussion underscoring the possibility of combining visual and econometric inference. That inferences can complement one another is not unique to visual

[6]Because the maximum of the standard correlation measure depends on the marginal distributions of R and $R_{\hat{\theta}}$ (for example, if $\Pr(R=1) \neq \Pr(R_{\hat{\theta}}=1)$, then their correlation cannot be 1), which vary across $|d|$ and $\hat{\theta}$, correlation is difficult to interpret. Although we can follow the spirit of the proposals by Cohen (1960) and Davenport and El-Sanhurry (1991) and present correlations scaled relative to their maximum as determined by the marginal distributions, we choose not to do so for brevity. Instead, we only show one-sided p-values from Fisher's exact test—two-sided p-values are method-dependent because of the ambiguity in classifying contingency tables as extreme in the opposite direction (see for example Agresti, 1992).

inference, and it harkens back to the introduction of Fisher's method (Fisher, 1925 and Mosteller and Fisher, 1948) which combines independent statistical tests by summing the logarithms of their p-values. Although we do not explicitly ask our participants to state a p-value, we can impute it from their visual inference decisions. For example, we can follow Habiger and Peña (2011): for each $R_i = 1$ when seeing a graph that is generated with parameter γ and discontinuity d, we uniformly randomly draw a p-value, p_i, from the interval $[0, \hat{p}(\gamma, d)]$; for each $R_i = 0$, draw p_i from $[\hat{p}(\gamma, d), 1]$. We can then combine the individual p-values along with that from econometric inference via the proposed method by Wilson (2019), which allows the p-values to be dependent unlike in Fisher's method.

To see that using the joint distribution of R and $R_{\hat{\theta}}$ can refine the inference on the discontinuity/kink d, it is easiest to use a Bayesian approach for illustration. Suppose researchers have an ex-ante sense of the discontinuity magnitude in an empirical application based on their read of the literature, which is described by prior density $\psi(\cdot)$. They then analyze the data and conduct visual and econometric inferences and reject no-discontinuity in both cases, i.e. $R = 1$ and $R_{\hat{\theta}} = 1$. The posterior probability that the discontinuity magnitude falls in a given set A is

$$\frac{\int_{z \in A} \Pr(R = 1, R_{\hat{\theta}} = 1 | z) d\psi(z)}{\int_{z} \Pr(R = 1, R_{\hat{\theta}} = 1 | z) d\psi(z)}.$$

When R and $R_{\hat{\theta}}$ always agree, i.e. they are perfectly correlated, knowing R does not change the posterior. As the association between R and $R_{\hat{\theta}}$ lessens, knowing R provides more information.

Although this is a conceptual demonstration, it does illustrate—through both frequentist and Bayesian lenses—that combining visual and econometric inference can be useful. This is true especially when the two inference decisions are not strongly associated, which is generally the case for RKD and when there is no discontinuity in RDD as shown in Section F.

G Eyetracking Study: Details and Additional Analyses

As outlined in Section 3.5 and Appendix C.3, the eyetracking study consists of a fixed sequence of 11 graphs which we show to both expert and non-expert participants on a Tobii XL60 eyetracking terminal that records the location of the subject's gaze every 17 milliseconds (60 times per second).

Figure A.45 is a visual representation of the eyetracking microdata from one participant looking at a particular RD graph. It plots a sequence of fixations in connected circles against the RD graph seen by the participant. The number inside each circle represents the order in which the fixations took place. In

total, this participant fixated on 22 points in the graph over approximately 9 seconds. The time lengths vary across the fixations and are not represented in the figure. The initial fixation occurred at the upper right corner of this particular graph for every participant, which was induced by the placement of the cross on the previous screen (see Figure A.37). We therefore remove the first two fixations from the microdata for every participant-graph combination. We also remove all saccades before the third fixation. Further, we remove all eye gazes that lie outside the outermost axis ticks because they mostly coincide with axis labels.

In the resulting data set, we transform the eyegaze pixel coordinate data into units of the (x, y) variables of each graph and compute a set of summary measures from the eyetracking microdata. We present the means of these measures in Table A.10. As mentioned in Section 4.5, the experts perform better than the non-experts. The 15 percentage point expert-non-expert success gap translates to 1.7 graphs out of the 11 in total. As seen in the second and third rows of the table, experts spend more than twice as long on each graph and have more than twice as many fixations than non-experts. Next, we attempt to measure the "visual bandwidth" by calculating the horizontal range and horizontal standard deviation of the eyegaze coordinates. Out of a maximum horizontal range of 2 (recall that the support of the x-variable is normalized to be between -1 and 1), the overall average is 1.19, with the experts covering a 30% longer distance. The overall average standard deviation of the eyegazes is much smaller at 0.24, and again we see experts using a larger ocular bandwidth.

The next set of measures assesses the extent to which participants visually trace out the conditional expectation function. First, we compute the square root of the eyegaze IMSE, the RIMSE. The IMSE is the average squared deviation of the eyegaze from the true underlying CEF of the DGP, and we scale the RIMSE by σ. The average RIMSE is 0.25σ and shows little difference between experts and non-experts. Second, we fit RD regression models to the eyegaze data using CCT's `rdrobust` package and record the conventional and bias-corrected local linear estimates. We then compute the absolute deviation of these estimates from the true discontinuity in the DGP and, consistent with our discontinuity specification in Section 3.2, scale this deviation by σ. One complication is that `rdrobust` cannot compute the bandwidths when there are few distinct eyegaze points, and the resulting conventional and bias-corrected local estimates are both missing. Not surprisingly, there are more missing estimates for non-experts (13%, or 13 out of 99 graphs) than experts (2%, or 1 out of 66 graphs) given that experts spend more time on each graph and exhibit a larger number of fixations. In these instances, we replace the conventional and bias-corrected estimates by

global linear and global quadratic estimates, respectively.[7] The average deviation is 0.23σ for conventional estimates and 0.24σ for bias-corrected estimates. The average deviation is slightly lower among experts for conventional estimates and more so for bias-corrected estimates. However, this difference is driven by the global regression estimates—within the 151 graphs for which local RD estimates can be computed by `rdrobust`, the difference in the deviations between experts and non-experts virtually disappears.

We test for differences in the summary measures between experts and non-experts, accounting for the potential covariance for the same individual across graphs by clustering at the participant level. Because of the small number of participants, we implement the test via wild bootstrap (see Cameron and Miller, 2015 for a review). As highlighted in Section 4.5, the only measures that are significantly different between the experts and non-experts at the 5% level are time spent on each graph and whether the local RD eyegaze estimates are available. The difference in one other eyegaze measure, horizontal range, is statistically significant at the 10% level, and so is the difference in the outcome variable, correct discontinuity classification.

For our prediction exercises, we split the data into a training sample and a testing sample, use the training sample to train the models, and compute the out-of-sample prediction error in the testing sample.[8] To gauge the robustness of the prediction results, we use two schemes to split our training and testing samples. In scheme A, we randomly select data from two-thirds of the 165 graphs to be the training sample. In scheme B, we randomly select all data from two-thirds of the participants (six of the nine non-experts and four of the six experts) to be the training sample. In both schemes, the out-of-sample prediction error rates range between 25% and 40%. In scheme A, the respective error rates are 25%, 40%, 33%, and 33% for penalized logit, pruned tree, bagging, and random forest. They are 31%, 29%, 33% and 29% in scheme B.

The penalized logit coefficient path plots in Figure A.46 trace out the coefficients on the predictors as functions of the penalty parameter (Lambda). In scheme A, the first nine predictors to turn on (become nonzero) are eight of the graph identifiers and the expert indicator. In scheme B, the graph identifiers are also among the first to turn on, except the coefficients on RIMSE and the absolute deviation of the bias-corrected RD estimate also become nonzero at high values of the penalty parameter. In fact, both coefficients are negative, indicating that higher RIMSE and absolute deviation of eyegaze RD estimates are associated with a lower probability of successful inference, which makes intuitive sense. However, these are not robust

[7] The bias-corrected local linear estimate obtained by using the same main and pilot bandwidth from Calonico et al. (2014) is numerically equivalent to the local quadratic estimate with the same bandwidth. Therefore, we can alternatively think of these global regressions as producing conventional and bias-corrected local linear estimates by using a large bandwidth.

[8] We select the tuning parameter in penalized-logit and pruned tree by five-fold cross validation.

associations: in scheme A, the RIMSE coefficient is instead positive, and although the coefficient on the absolute deviation of the bias-corrected estimate is still negative, the model selects this variable only when the penalty parameter is low, and the variable appears to counteract the *positive* coefficient on the absolute deviation of the conventional estimate. We see this offsetting pattern of the conventional and bias-corrected RD estimates in scheme B as well. Analogously, the horizontal eyegaze standard deviation and horizontal range also have opposite signs from the penalized logit model, and it is unclear whether a larger or smaller visual bandwidth is conducive to successful inference.

The tree-based methods paint a broadly similar picture on the importance of the various predictors as penalized logit. In both training-testing schemes, the pruned tree is a stump and only splits by the Graph1 indicator (graph numbers also correspond to DGP numbers, i.e. Graph1 is from RD DGP1). Figure A.47 presents the importance plots from bagging and random forest models, where each variable's importance is measured by the decrease in prediction accuracy if the variable is omitted from the model (see Friedman, Hastie, and Tibshirani, 2001 for details). The relative importance of the eyegaze measures varies by method and scheme, but the Graph1 and Graph4 indicators consistently stand out.

We also try re-running the prediction exercise after removing the four graphs that are "too easy" or "too hard," i.e. for which at least 14 out of 15 participants classify discontinuity correctly/incorrectly (Graph 1 is too hard, while Graphs 4, 10 and 11 are too easy). The resulting prediction error rates increase to between 40% and 50%, but the most important predictors still appear to be several graph identifiers. Our visual attention summary measures continue to lack robust predictive potency within the remaining seven graphs.

Based on these findings, we cannot pinpoint ocular patterns that robustly predict successful visual inference. We conclude that performance heterogeneity in visual inference is likely driven by the processing of visual signals and not by their acquisition as quantified by our eyegaze summary measures.

Appendix References

BUGNI, F. A., I. A. CANAY, AND A. M. SHAIKH (2019): "Inference Under Covariate-Adaptive Randomization with Multiple Treatments," *Quantitative Economics*, 10, 1747–1785.

CALONICO, S., M. D. CATTANEO, AND M. H. FARRELL (2020): "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs," *The Econometrics Journal*, 23, 192–210.

CALONICO, S., M. D. CATTANEO, AND R. TITIUNIK (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

——— (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110, 1753–1769.

CAMERON, A. C. AND D. L. MILLER (2015): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372.

CARD, D., D. S. LEE, AND Z. PEI (2009): "Quasi-Experimental Identification and Estimation in the Regression Kink Design," Princeton University Industrial Relations Section Working Paper 553.

CHAPMAN, J., E. SNOWBERG, S. WANG, AND C. CAMERER (2018): "Loss Attitudes in the US Population: Evidence from Dynamically Optimized Sequential Experimentation (DOSE)," National Bureau of Economic Research Working Paper 25072.

DURRETT, R. (2010): *Probability: Theory and Examples*, Cambridge University Press, 4th ed.

FISHER, R. A. (1925): *Statistical Methods for Research Workers*, Oliver and Boyd (Edinburgh).

FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2001): *The Elements of Statistical Learning*, vol. 1, Springer Series in Statistics New York.

HABIGER, J. D. AND E. A. PEÑA (2011): "Randomised P-Values and Nonparametric Procedures in Multiple Testing," *Journal of Nonparametric Statistics*, 23, 583–604.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design," National Bureau of Economic Research Working Paper 7131.

HOEFFDING, W. (1956): "On the Distribution of the Number of Successes in Independent Trials," *Annals of Mathematical Statistics*, 27, 713–721.

IMBENS, G. AND K. KALYANARAMAN (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of economic studies*, 79, 933–959.

KAHNEMAN, D. AND A. TVERSKY (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

LEE, D. S. (2008): "Randomized Experiments from Non-Random Selection in US House Elections," *Journal of Econometrics*, 142, 675–697.

MCCRARY, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698–714.

MOSTELLER, F. AND R. A. FISHER (1948): "Questions and Answers," *The American Statistician*, 2, 30–31.

NICKELL, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica: Journal of the Econometric Society*, 49(6), 1417–1426.

PEI, Z., J.-S. PISCHKE, AND H. SCHWANDT (2019): "Poorly Measured Confounders are More Useful on the Left than on the Right," *Journal of Business & Economic Statistics*, 37, 205–216.

PEI, Z. AND Y. SHEN (2017): "The Devil is in the Tails: Regression Discontinuity Design with Measurement Error in the Assignment Variable," in *Regression Discontinuity Designs: Theory and Applications*, ed. by M. D. Cattaneo and J. C. Escanciano, vol. 38 of *Advances in Econometrics*, 455–502.

PERCUS, O. E. AND J. K. PERCUS (1985): "Probability Bounds on the Sum of Independent Nonidentiacally Distributed Binomial Random Variables," *SIAM Journal of Applied Mathematics*, 45, 621–640.

WILSON, D. J. (2019): "The Harmonic Mean p-Value for Combining Dependent Tests," *Proceedings of the National Academy of Sciences*, 116(4), 1195–1200.

Appendix Tables

Table A.1: Data-Driven Global Polynomial Orders for Fit Lines

DGP	AIC	BIC	Lee and Lemieux Dummy F-Test
DGP1	5	0	0
DGP2	0	0	0
DGP3	5	5	0
DGP4	5	1	0
DGP5	0	0	0
DGP6	5	5	3
DGP7	2	2	0
DGP8	4	1	3
DGP9	0	0	0
DGP10	2	2	1
DGP11	3	0	0

Notes: Table presents optimal polynomial orders based on three algorithms: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the procedure from Lee and Lemieux (2010) based on a joint F-test of significance for the bin dummies. We restrict all algorithms to polynomial orders zero through five.

Table A.2: Timeline of RDD Experiments

Phase	Holding fixed	Treatments	Date	# participants recruited	# completions
RDD1	bin spacing: ES fitlines: no vertical line: yes	binwidth: large vs small # axis scaling: default vs 2x default	November 13-16, 2018	4*88=352	330 (94%)
RDD2	scaling: default fitlines: no vertical line: yes	binwidth: large vs small # bin spacing: ES vs QS	February 11-12, 2019	4*88=352	325 (92%)
RDD3	binwidth: small bin spacing: ES scaling: default	fitlines: no; center line: yes fitlines: no; center line: no fitlines: yes; center line: yes	February 27, 2019	3*88=264	248 (94%)
RDD4	bin spacing: ES scaling: default vertical line: yes	binwidth: large vs small # fitlines: yes vs no	October 28-29, 2019	4*88=352	340 (97%)

Table A.3: Timeline of RKD Experiments

Phase	Holding fixed	Treatments (between subject)	Treatments (within subject)	Date	# participants recruited	# completions
Pilot	bin spacing: QS fit lines: no vertical line: yes scaling: default	binwidth: large vs small # detrending: yes vs no	derivative vs level	April 24-25, 2019	4*72=288	282 (98%)
RKD	bin spacing: QS fit lines: no vertical line: yes	binwidth: large vs small # detrending: yes vs no	derivative vs level	June 26-28, 2019	4*72=288	250 (87%)

Table A.4: Treatment Effects: Phase RDD1

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; Default y-axis	0.197	0.209	0.177	0.140	-0.024	-0.012
	(0.039)	(0.045)	(0.042)	(0.046)	(0.030)	(0.019)
Small bin; Large y-axis	-0.024	-0.018	-0.056	0.008	-0.034	-0.035
	(0.029)	(0.043)	(0.042)	(0.043)	(0.029)	(0.026)
Large bin; Large y-axis	0.142	0.186	0.120	0.093	0.013	-0.001
	(0.036)	(0.040)	(0.042)	(0.042)	(0.025)	(0.017)
Small bin; Default y-axis (Mean)	0.054	0.181	0.367	0.669	0.922	0.988
Number of graphs	660	660	660	660	660	330
Number of players	330	330	330	330	330	330

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni, Canay, and Shaikh (2019) for stratified experiments, which are asymptotically exact.

Table A.5: Treatment Effects: Phase RDD2

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; Equal spacing	0.256	0.207	0.204	0.155	0.048	0.011
	(0.042)	(0.039)	(0.045)	(0.039)	(0.033)	(0.029)
Small bin; Quantile spacing	0.034	0.009	-0.012	-0.008	-0.030	-0.013
	(0.033)	(0.039)	(0.046)	(0.046)	(0.036)	(0.033)
Large bin; Qunatile spacing	0.157	0.185	0.217	0.138	0.030	0.024
	(0.038)	(0.040)	(0.046)	(0.039)	(0.031)	(0.023)
Small bin; Equal spacing (Mean)	0.054	0.169	0.355	0.657	0.892	0.964
Number of graphs	650	650	650	650	650	325
Number of players	325	325	325	325	325	325

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni et al. (2019) for stratified experiments, which are asymptotically exact.

Table A.6: Treatment Effects: Phase RDD3

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Fit line; Vertical line	0.142	0.215	0.172	0.051	0.057	-0.002
	(0.038)	(0.046)	(0.044)	(0.045)	(0.029)	(0.025)
No fit line; No vertical line	0.017	0.012	-0.007	-0.052	-0.026	-0.023
	(0.024)	(0.043)	(0.045)	(0.043)	(0.031)	(0.025)
No fit line; Vertical line (Mean)	0.037	0.159	0.348	0.634	0.860	0.976
Number of graphs	496	496	496	496	496	248
Number of players	248	248	248	248	248	248

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni et al. (2019) for stratified experiments, which are asymptotically exact.

Table A.7: Treatment Effects: Phase RDD4

	Dependent variable: player reports discontinuity					
	True discontinuity magnitude =					
	0	0.1944 σ	0.324 σ	0.54 σ	0.9 σ	1.5 σ
Large bin; No fit line	0.145	0.283	0.223	0.239	0.074	0.027
	(0.038)	(0.044)	(0.047)	(0.042)	(0.033)	(0.030)
Small bin; Fit line	-0.019	0.088	0.047	0.059	0.024	0.002
	(0.030)	(0.042)	(0.049)	(0.046)	(0.036)	(0.034)
Large bin; Fit line	0.230	0.300	0.239	0.207	0.058	0.003
	(0.045)	(0.046)	(0.047)	(0.044)	(0.034)	(0.034)
Small bin; No fit line (Mean)	0.074	0.154	0.395	0.611	0.870	0.951
Number of graphs	680	680	680	680	680	340
Number of players	340	340	340	340	340	340

Notes: We have a stratified randomized experiment, where each of the 11 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni et al. (2019) for stratified experiments, which are asymptotically exact.

Table A.8: Treatment Effects: Phase RKD (Levels)

	Dependent variable: player reports kink				
	True kink magnitude =				
	0	0.1944 ξ	0.324 ξ	0.54 ξ	0.9 ξ
Raw; Big bin	-0.008	0.054	0.107	-0.003	-0.084
	(0.040)	(0.046)	(0.052)	(0.057)	(0.064)
Residual; Small bin	0.013	0.002	0.046	0.111	-0.031
	(0.041)	(0.048)	(0.057)	(0.053)	(0.058)
Residual; Big bin	0.018	0.242	0.291	0.175	-0.015
	(0.041)	(0.051)	(0.055)	(0.054)	(0.059)
Raw; Small bin (Mean)	0.147	0.248	0.352	0.543	0.827
Number of graphs	614	608	607	614	307
Number of players	250	250	250	250	250

Notes: We have a stratified randomized experiment, where each of the 9 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni et al. (2019) for stratified experiments, which are asymptotically exact.

Table A.9: Treatment Effects: Phase RKD (Derivatives)

	Dependent variable: player reports kink				
	True kink magnitude =				
	0	0.1944 ξ	0.324 ξ	0.54 ξ	0.9 ξ
Derivative; Big bin	0.143	-0.004	0.089	0.153	0.302
	(0.040)	(0.041)	(0.045)	(0.042)	(0.058)
Derivative; Small bin (Mean)	0.235	0.312	0.294	0.306	0.335
Number of players	250	250	250	250	250
Number of graphs	614	608	607	614	307

Notes: We have a stratified randomized experiment, where each of the 9 strata is determined by the DGPs seen for every discontinuity magnitude. The treatment effect estimates come from regressing the participants' responses on treatment indicators and strata fixed effects. We obtain standard errors using the procedure from Bugni et al. (2019) for stratified experiments, which are asymptotically exact.

Table A.10: Summary Measures of Eyegaze Data

	Average across graphs			
	(1)	(2)	(3)	(4)
	Experts	Non-experts	Overall	p-value of difference between (1) and (2)
Number of Correct Classifications	0.68	0.53	0.59	0.078
Time Spent per Graph in Seconds	16.8	6.9	10.9	0.046
Number of Fixations per Graph	45.8	20.5	30.6	0.112
Horizontal Range of Eyegazes	1.40	1.05	1.19	0.090
Horizontal Standard Deviation of Eyegazes	0.26	0.22	0.24	0.268
RIMSE of Eyegazes (scaled)	0.25	0.26	0.25	0.615
Conventional Local Linear RD Estimate from Eyegazes				
Absolute Deviation from True Discontinuity (scaled)	0.22	0.23	0.23	0.136
Discontinuity Estimate Unavailable	0.02	0.13	0.08	0.032
Bias-corrected Local Linear RD Estimate from Eyegazes				
Absolute Deviation from True Discontinuity: Bias-corrected RD Estimate	0.22	0.25	0.24	0.103
Discontinuity Estimate Unavailable	0.02	0.13	0.08	0.032
Number of Participants	6	9	15	
Number of Graphs	66	99	165	

Notes: The statistical test in column (4) accounts for correlation across the same participant via clustering, and we use wild bootstrap due to the small number of clusters.

Table A.12: Covariate Means and Balance across Treatment Arms: Phase RDD1

	T1	T2	T3	T4	p-value
Female	0.687	0.738	0.759	0.688	0.649
Completed college	0.518	0.512	0.482	0.625	0.268
Age 18 to 22	0.422	0.571	0.554	0.425	0.088
Age 23 to 49	0.494	0.381	0.398	0.487	0.318
Age 50 or older	0.060	0.048	0.048	0.087	0.742
Graduate stats knowledge	0.145	0.119	0.133	0.050	0.091
Passed attention check	0.867	0.857	0.843	0.912	0.516
Attrition rate	0.057	0.045	0.057	0.091	0.690
Participants	83	84	83	80	

Notes: The joint p-value for the significance of the covariates is 0.175.

Table A.11: Timeline of Expert Study

Participant pool	Binwidth choice in Part 1	Date	# participants recruited	# completions
Cornell Econometrics Workshop (RD only)	RD: small binwidth, even spacing	May 7, 2019	-	12
UC Irvine Econometrics Series (RD only)	RD: small binwidth, even spacing	May 21, 2019	-	15
Princeton Quantitative Social Science Colloquium (RD only)	RD: large binwidth, even spacing	October 11, 2019	-	48
NBER + IZA (RD only, Pilot)	RD: small binwidth, even spacing	July 28 - August 2, 2019	101	7 (7%)
NBER + IZA Email list (RD and RK, Pilot)	RD: small binwidth, even spacing RK: big binwidth, quantile spacing	August 12 - 19, 2019	200	21 (11%)
NBER + IZA Email list (RD and RK)	RD: small binwidth, even spacing RK: big binwidth, quantile spacing	August 26 - September 3, 2019	1,000	94 (9%)

Table A.13: Covariate Means and Balance across Treatment Arms: Phase RDD2

	T1	T2	T3	T4	p-value
Female	0.614	0.623	0.617	0.726	0.325
Completed college	0.494	0.429	0.543	0.512	0.526
Age 18 to 22	0.518	0.584	0.481	0.488	0.548
Age 23 to 49	0.386	0.364	0.481	0.429	0.447
Age 50 or older	0.084	0.039	0.037	0.071	0.485
Graduate stats knowledge	0.072	0.143	0.099	0.119	0.499
Passed attention check	0.928	0.870	0.877	0.940	0.306
First time player	0.711	0.805	0.802	0.762	0.466
Attrition rate	0.057	0.125	0.080	0.045	0.264
Participants	83	77	81	84	

Notes: The joint p-value for the significance of the covariates is 0.496.

Table A.14: Covariate Means and Balance across Treatment Arms: Phase RDD3

	T1	T2	T3	p-value
Female	0.683	0.725	0.616	0.323
Completed college	0.659	0.650	0.558	0.339
Age 18 to 22	0.439	0.463	0.523	0.529
Age 23 to 49	0.488	0.512	0.419	0.449
Age 50 or older	0.073	0.025	0.047	0.345
Graduate stats knowledge	0.159	0.113	0.070	0.184
Passed attention check	0.902	0.900	0.907	0.988
First time player	0.488	0.500	0.453	0.823
Attrition rate	0.068	0.091	0.023	0.087
Participants	82	80	86	

Notes: The joint p-value for the significance of the covariates is 0.653.

Table A.15: Covariate Balance across Treatment Arms: Phase RDD4

	T1	T2	T3	T4	p-value
Female	0.654	0.782	0.733	0.744	0.327
Completed college	0.543	0.655	0.442	0.535	0.040
Age 18 to 22	0.568	0.471	0.663	0.570	0.084
Age 23 to 49	0.420	0.506	0.291	0.407	0.031
Age 50 or older	0.012	0.023	0.035	0.023	0.797
Graduate stats knowledge	0.160	0.115	0.070	0.047	0.067
Passed attention check	0.914	0.897	0.953	0.907	0.427
First time player	0.556	0.563	0.628	0.581	0.771
Attrition rate	0.080	0.011	0.023	0.023	0.185
Participants	81	87	86	86	

Notes: The joint p-value for the significance of the covariates is 0.1365.

Table A.16: Covariate Means and Balance across Treatment Arms: Phase RKD

	T1	T2	T3	T4	p-value
Female	0.758	0.721	0.705	0.710	0.903
Completed college	0.515	0.492	0.541	0.500	0.952
Age 18 to 22	0.667	0.574	0.557	0.597	0.588
Age 23 to 49	0.288	0.361	0.393	0.387	0.554
Age 50 or older	0.030	0.066	0.016	0.016	0.522
Graduate stats knowledge	0.030	0.115	0.131	0.113	0.052
Passed attention check	0.909	0.967	0.951	0.919	0.466
First time player	0.455	0.410	0.557	0.468	0.417
Attrition rate	0.250	0.307	0.307	0.295	0.804
Participants	66	61	61	62	

Notes: The joint p-value for the significance of the covariates is 0.545.

Table A.17: Correlations of Correct Classification with Subject Covariates for RD

	Phase I	Phase II	Phase III	Phase IV
Female	-0.001	-0.007	-0.019	0.017
	(0.014)	(0.014)	(0.018)	(0.017)
Completed college	0.004	-0.018	-0.025	0.015
	(0.019)	(0.018)	(0.019)	(0.023)
Age 23 to 49	-0.006	0.003	-0.012	0.009
	(0.021)	(0.019)	(0.021)	(0.023)
Age 50 or older	0.004	0.037	0.094	0.055
	(0.026)	(0.031)	(0.033)	(0.073)
Graduate stats knowledge	-0.013	0.030	0.053	-0.048
	(0.022)	(0.022)	(0.028)	(0.029)
Passed attention check	0.055	0.043	-0.014	-0.011
	(0.023)	(0.026)	(0.029)	(0.031)
First time player	NA	0.000	-0.001	0.018
	(NA)	(0.016)	(0.017)	(0.015)
Joint test p-value	0.321	0.460	0.009	0.389

Notes: This table presents the coefficients from regressing whether study participants are correct in classifying an RD graph on their characteristics. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table A.18: Correlations of Correct Classification with Subject Covariates for RK

	RK Phase I
Female	-0.009
	(0.018)
Completed college	-0.030
	(0.021)
Age 23 to 49	-0.003
	(0.023)
Age 50 or older	-0.047
	(0.035)
Graduate stats knowledge	0.017
	(0.033)
Passed attention check	-0.013
	(0.039)
Joint test p-value	0.347

Notes: This table presents the coefficients from regressing whether study participants are correct in classifying an RK graph on their characteristics. Standard errors are clustered at the participant level.

Table A.19: Predictions of Power on DGP Characteristics

	Phase RDD1	Phase RDD2	Phase RDD3	Phase RDD4
LHS deriv-, RHS deriv-, discont-	0.244	0.262	0.267	0.308
	(0.032)	(0.033)	(0.039)	(0.031)
LHS deriv+, RHS deriv-, discont-	-0.005	0.029	-0.020	0.096
	(0.039)	(0.042)	(0.047)	(0.041)
LHS deriv-, RHS deriv+, discont-	0.024	0.103	0.173	0.102
	(0.058)	(0.057)	(0.068)	(0.054)
LHS deriv+, RHS deriv+, discont-	-0.111	-0.169	-0.103	-0.110
	(0.022)	(0.022)	(0.026)	(0.022)

Notes: This table presents the coefficients from regressions of whether the participant is correct in classifying discontinuous graphs on the interactions of the signs of left and right derivatives of the CEF with the sign of discontinuity. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table A.20: Predictions of Size on DGP Characteristics

	Phase RDD1	Phase RDD2	Phase RDD3	Phase RDD4
LHS deriv-, RHS deriv-	0.052	0.049	-0.003	0.093
	(0.035)	(0.037)	(0.029)	(0.036)
LHS deriv+, RHS deriv-	-0.040	-0.018	0.072	-0.048
	(0.032)	(0.035)	(0.037)	(0.032)
LHS deriv-, RHS deriv+	-0.124	-0.168	-0.032	-0.074
	(0.026)	(0.022)	(0.033)	(0.039)
LHS deriv+, RHS deriv+ (Mean)	0.141	0.168	0.077	0.154
	(0.020)	(0.022)	(0.018)	(0.020)

Notes: This table presents the coefficients from regressions of whether the participant is correct in classifying continuous graphs on the interactions of the signs of left and right derivatives of the CEF. Each of the four columns represents results from one of the four RD experiments. Standard errors are clustered at the participant level.

Table A.21: Dynamic Discontinuity Classification: AR(1) Regression

	(1)	(2)
Phase	Estimate of ρ from Eq. (A3)	Bias-Corrected Estimate of ρ
RDD1	-0.127	-0.041
	(0.038)	(0.046)
RDD2	-0.170	-0.088
	(0.035)	(0.039)
RDD3	-0.099	-0.009
	(0.029)	(0.036)
RDD4	-0.126	-0.034
	(0.035)	(0.041)
Expert	-0.205	-0.126
	(0.032)	(0.037)

Notes: Standard errors are in parentheses. Standard errors in both columns are clustered at the participant level, and those in column (2) additionally corrects the $O(1/S)$ bias from regression (A3) by inverting equation (18) of Nickell (1981).

Table A.22: Fisher's Exact Test of Association: p-values (One-Sided)

Panel A: Expert Visual vs Econometric Inferences (RDD)

	Estimator for Econometric Inference			
Discontinuity $\lvert d\rvert$	PQ	IK	CCT	AK
0	0.727	0.231	0.616	0.695
0.1944σ	0.000	0.000	0.000	0.000
0.324σ	0.000	0.000	0.000	0.000
0.54σ	0.000	0.000	0.001	0.000
0.9σ	0.073	.	0.042	0.112
1.5σ	.	.	.	.

Panel B: Expert Visual vs Econometric Inferences (RKD)

	Estimator for Econometric Inference				
Kink $\lvert d\rvert$	PQ	CCT1	CCT2	CCT3	CCT4
0	0.289	0.707	.	0.613	0.004
0.1944ξ	0.124	0.418	0.432	0.571	0.076
0.324ξ	0.456	0.212	0.600	0.080	0.194
0.54ξ	0.087	0.158	0.521	0.571	0.070
0.9ξ	0.191	0.572	0.626	0.258	0.601

Notes: Missing values indicate that a test always accepts or rejects the null hypothesis. The five columns correspond to the five estimators underlying the power functions in Figure 5. PQ is "Piecewise Quintic"; CCT1 is "CCT Robust"; CCT2 is "CCT Rob. Lin. w/ Reg. Tri."; CCT3 is "CCT Conv. Quad. Unif."; CCT4 is "CCT Conv. Lin. Unif.".

Table A.23: Timeline of Eyetracking Study

Participant pool	Date	# participants recruited	# completions
Nonexperts	July 15 & 17, 2019	11	10
Experts	July 30, 2019	6	6

Appendix Figures

Figure A.1: Same Data with Different Axis Scales for Correlation Detection

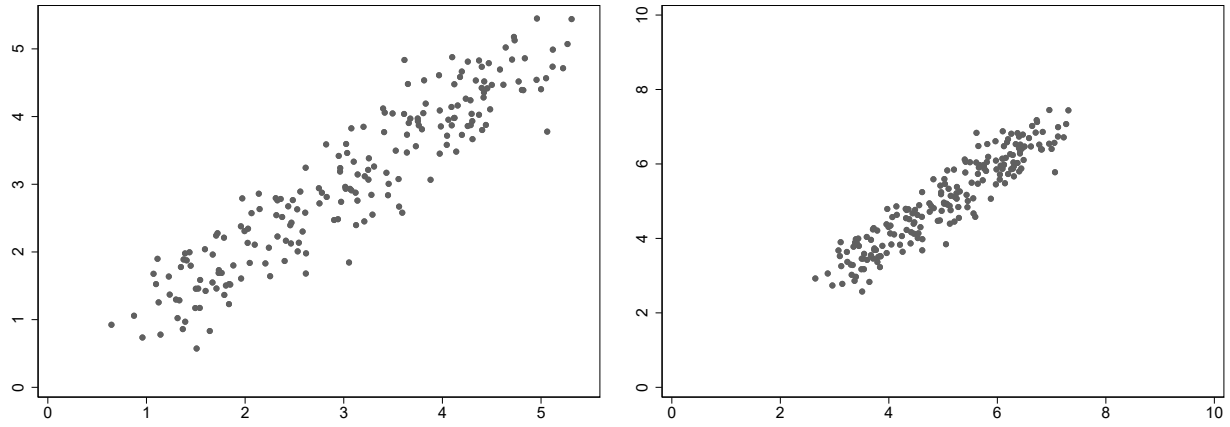


Figure A.2: Original Plot vs Plot with Residualized Y

(a) Original

(b) Residualized

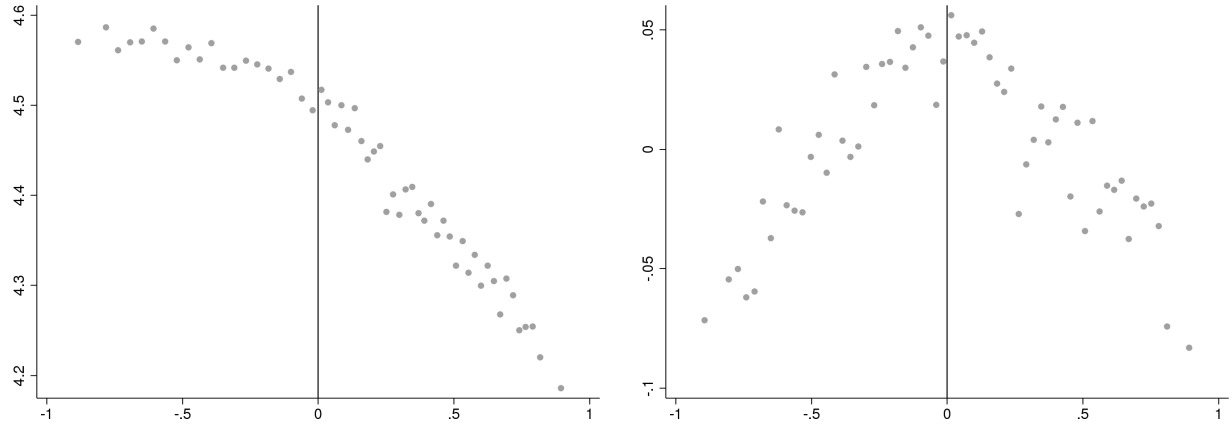
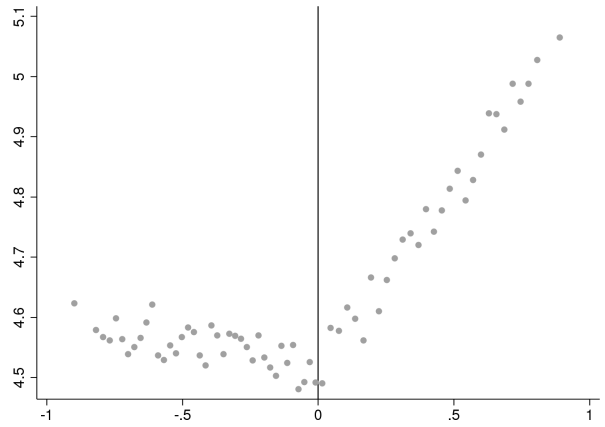


Figure A.3: Original Plot vs Plot of Estimated First Derivatives

(a) Original



(b) Estimated First Derivatives

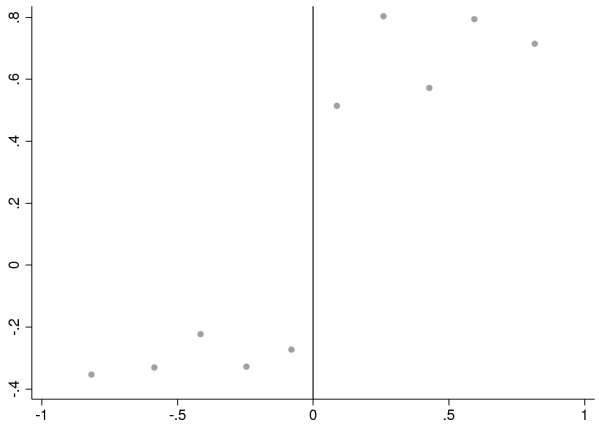
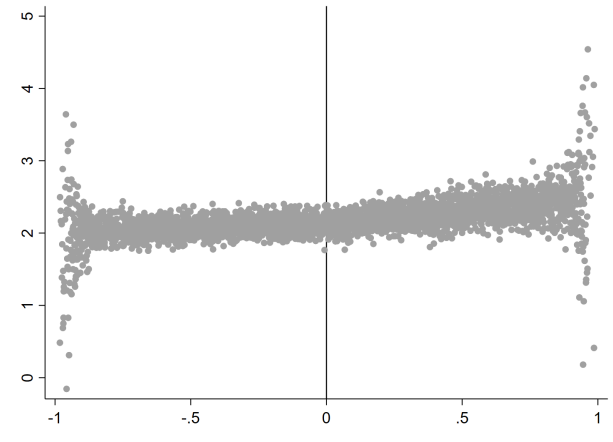


Figure A.4: RKD: Even Spacing vs Quantile Spacing Example

(a) Even Spacing



(b) Quantile Spacing

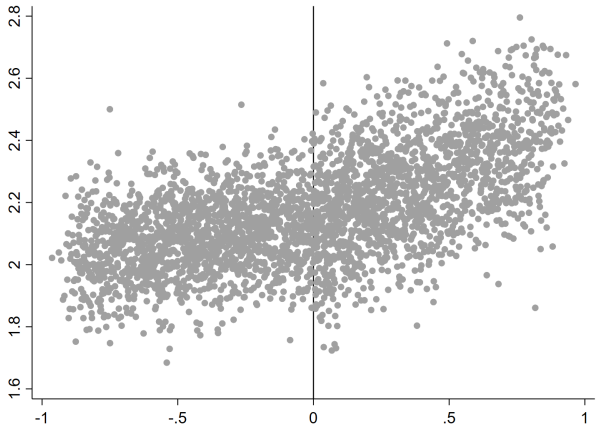


Figure A.5: RDD DGP Conditional Expectation Functions

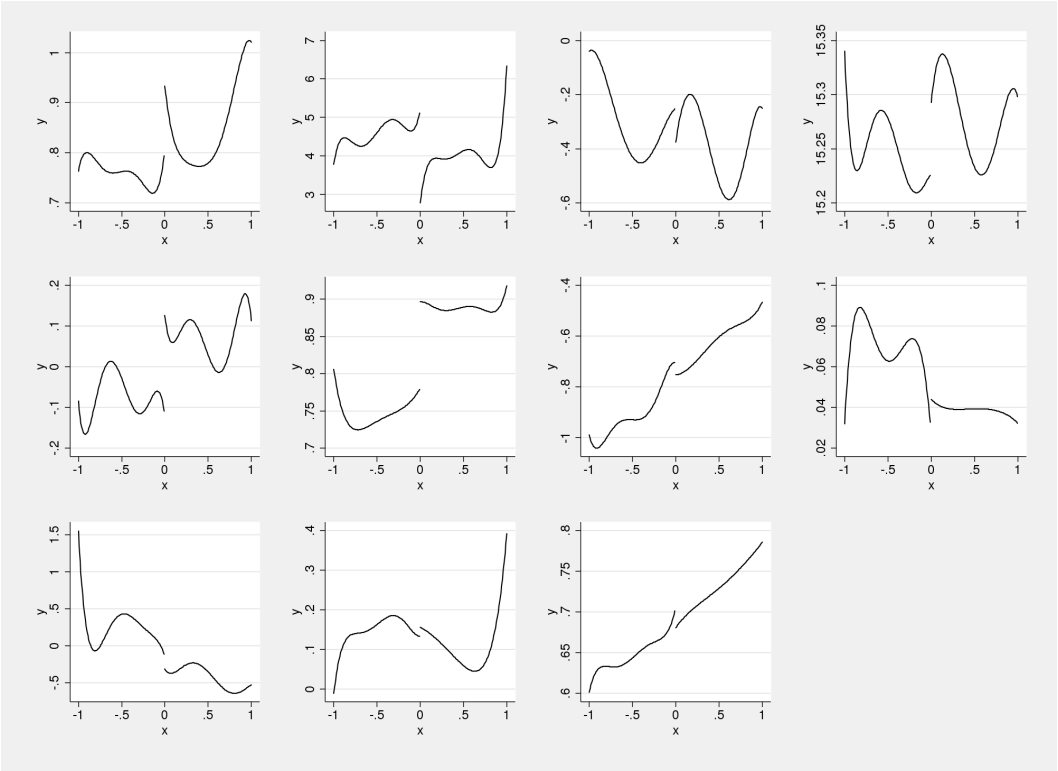
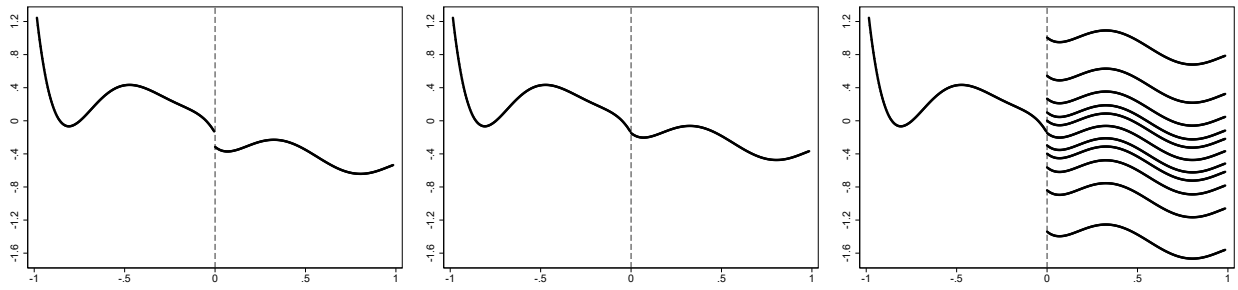


Figure A.6: Creation of RDD DGP Conditional Expectation Functions, DGP9



Notes: The leftmost figure plots the piecewise quintic CEF from the original microdata of the paper from which we calibrated DGP9. The central figure removes the discontinuity by setting the right intercept to equal the left intercept. The rightmost figure plots the final 11 CEFs for DGP9 corresponding to different levels of discontinuity by further changing the right intercept.

Figure A.7: RKD DGP Conditional Expectation Functions

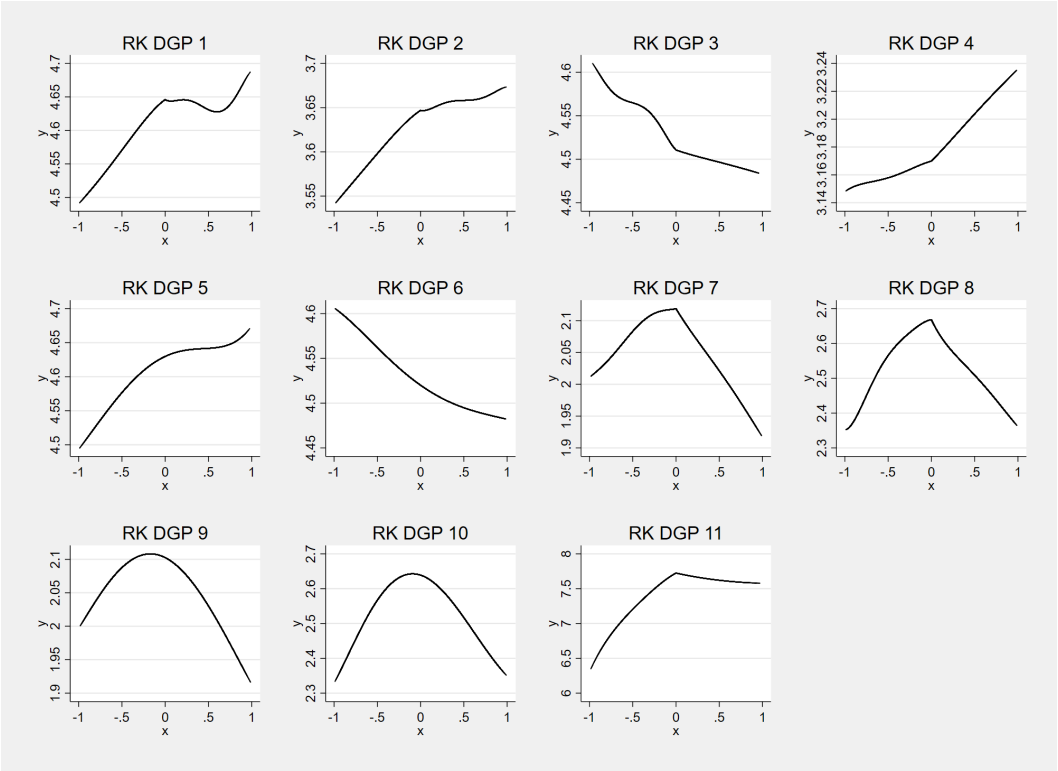
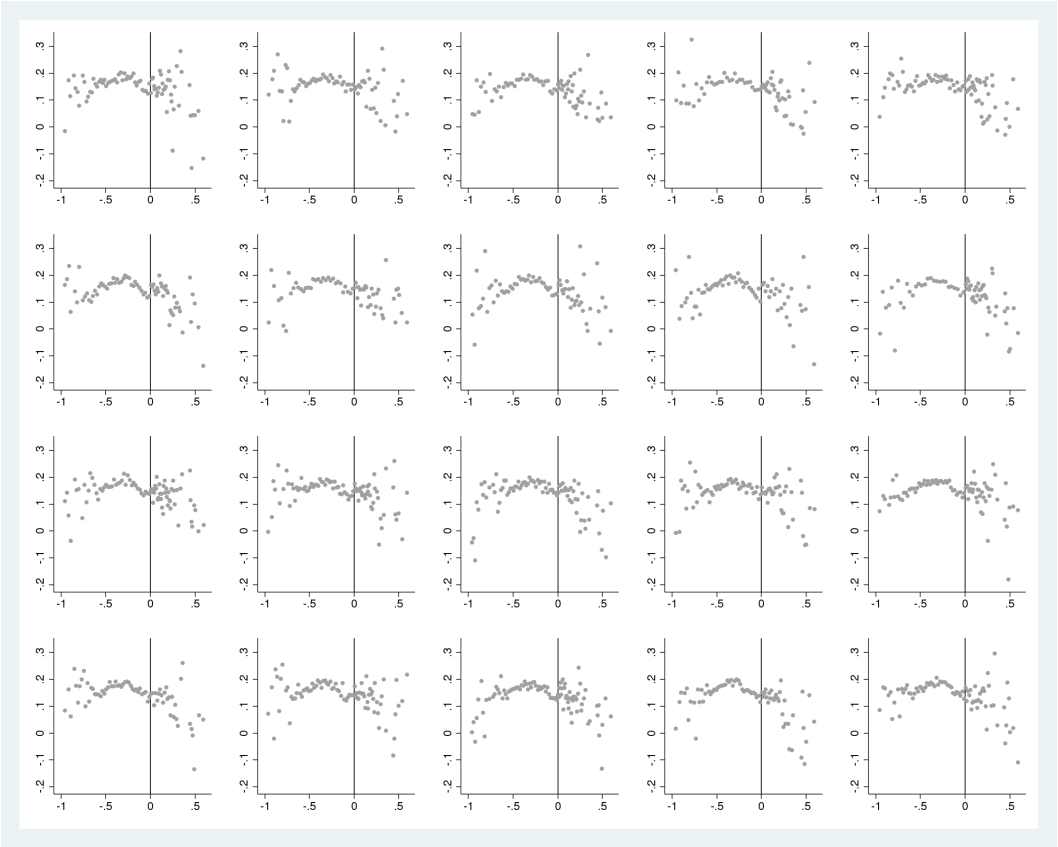
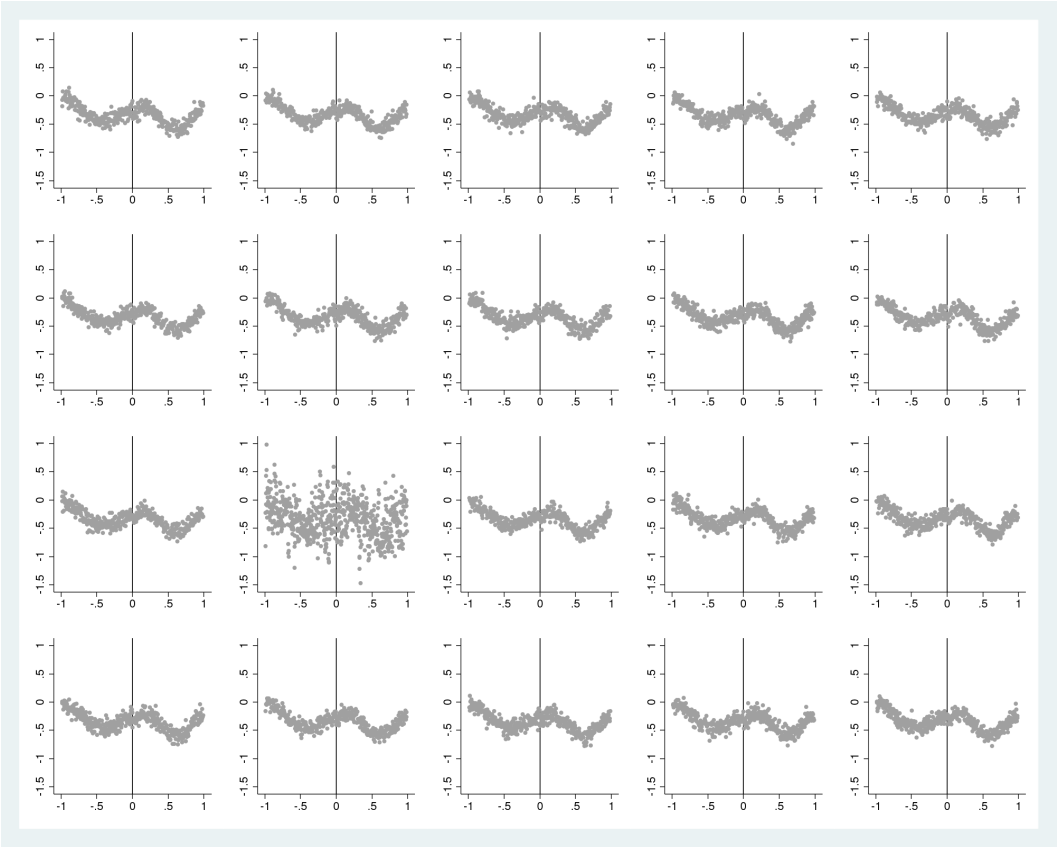


Figure A.8: Lineup Protocol Graph: RD DGP10



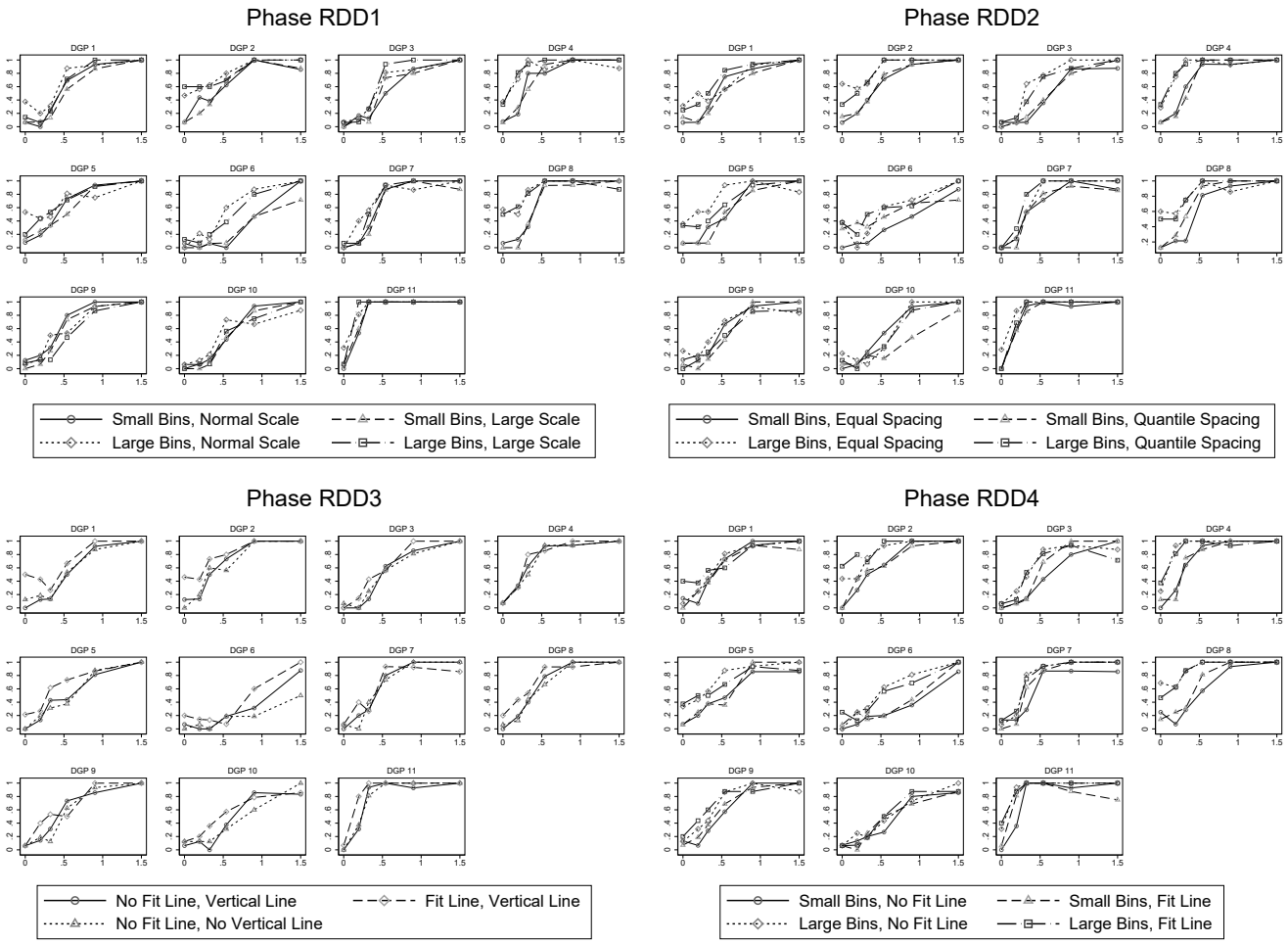
Notes: One of the 20 graphs is produced from the real data. The other 19 are produced from simulated data drawn from the DGP calibrated to the real data.

Figure A.9: Lineup Protocol Graph: RD DGP3



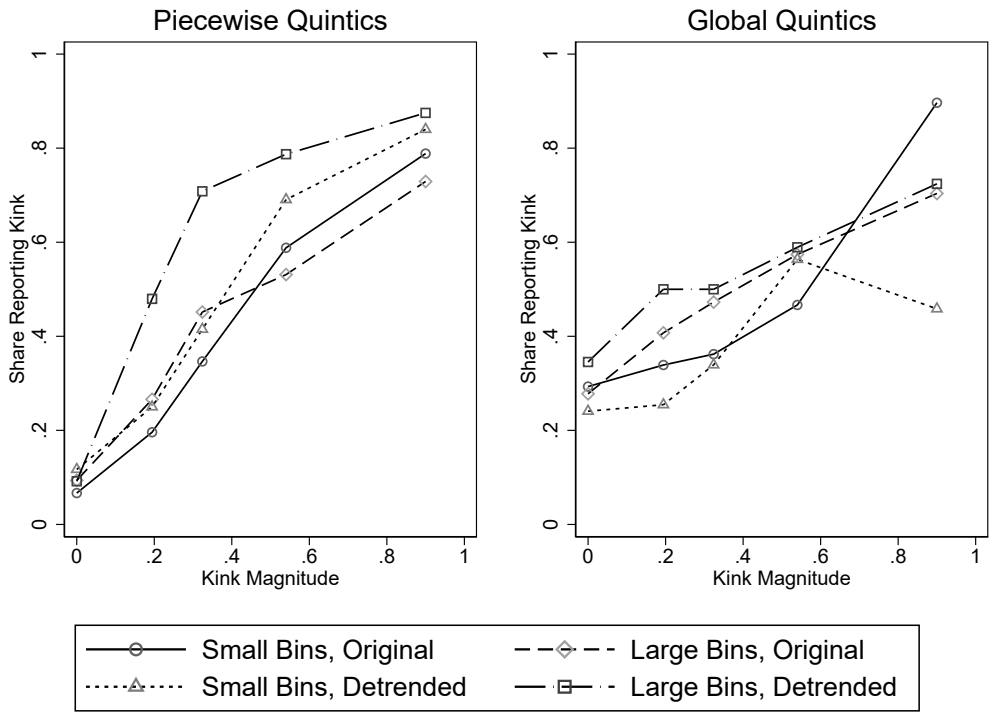
Notes: One of the 20 graphs is produced from the real data. The other 19 are produced from simulated data drawn from the DGP calibrated to the real data.

Figure A.10: RDD Phases: Power Functions by DGP



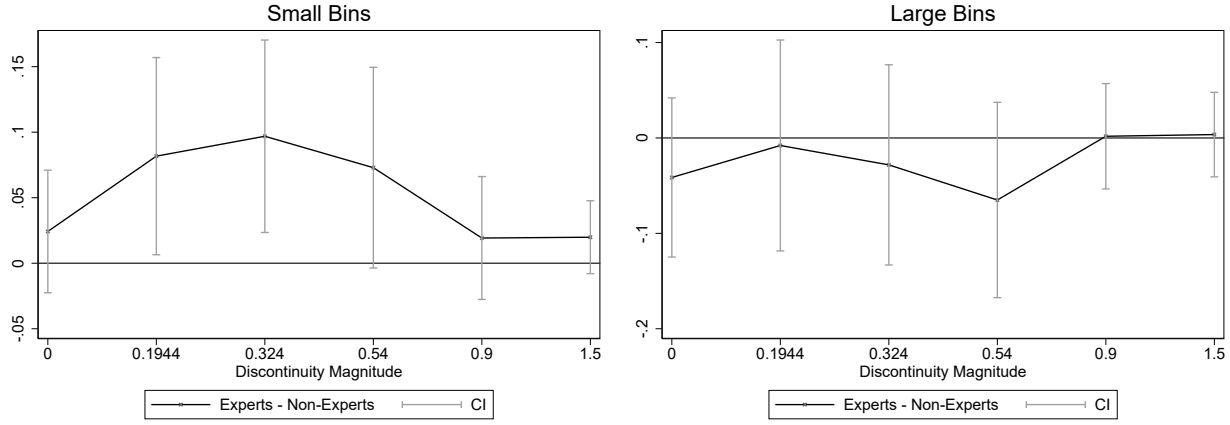
Notes: The figures here break up the power functions plotted in Figure 1 by DGP.

Figure A.11: RKD Phase: Power Functions By Quintic Type



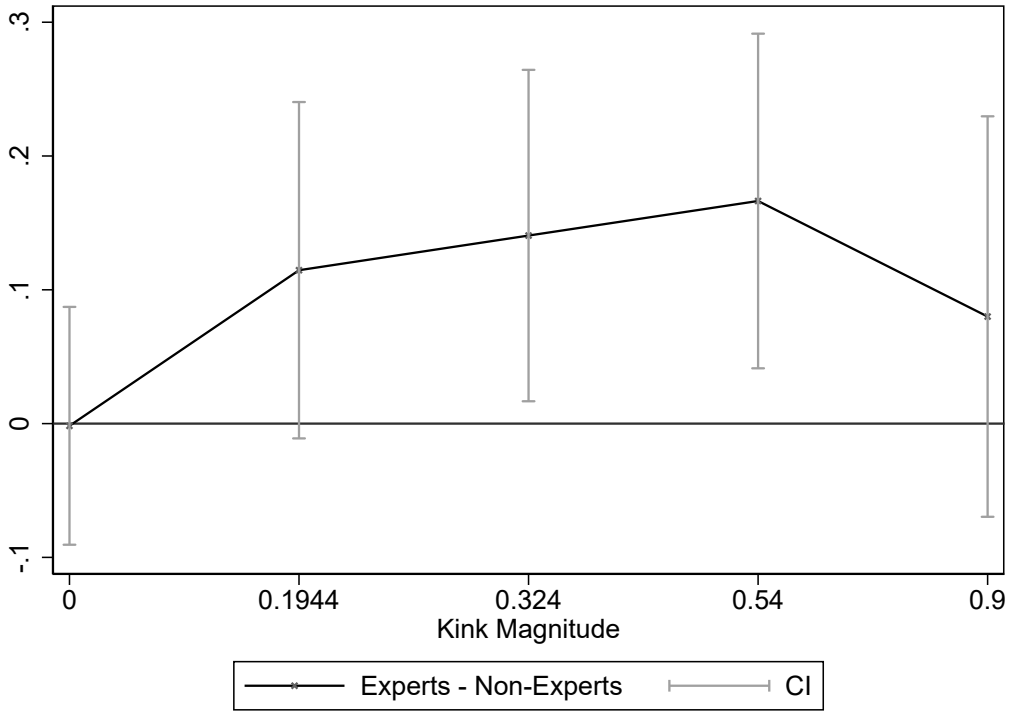
Notes: The figures here break up the power functions plotted in Figure 2 by two groups of DGPs—those specified with piecewise quintics and those specified with global quintics.

Figure A.12: RD: Expert vs Non-Expert Performance Differences



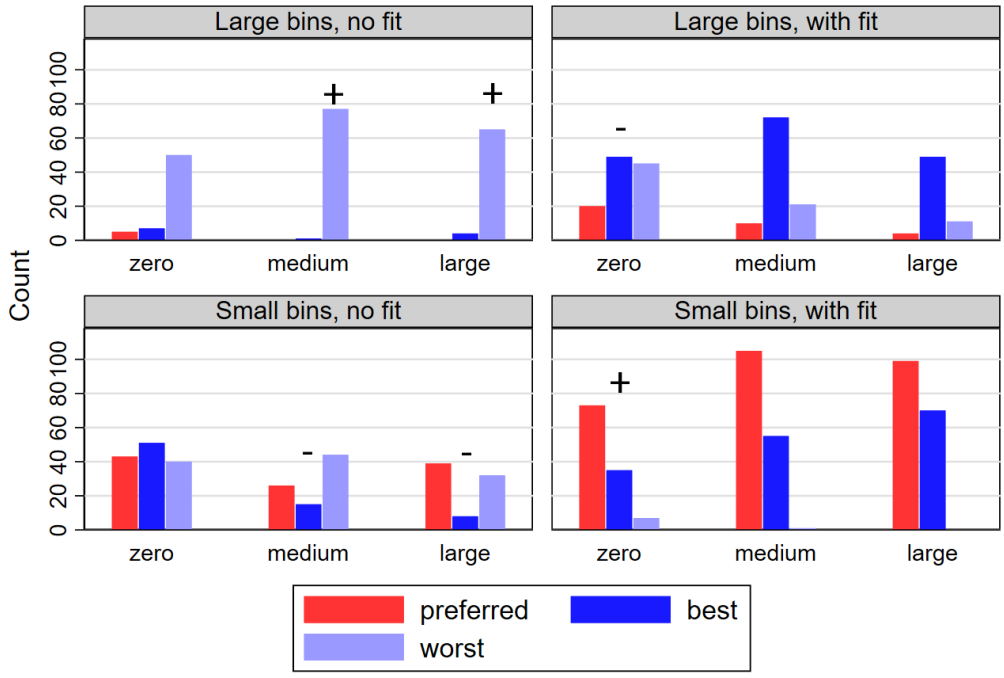
Notes: These figures plot the difference in the RD power functions between experts and non-experts from the left panel of Figure 3.

Figure A.13: RK: Expert vs Non-Expert Performance Differences



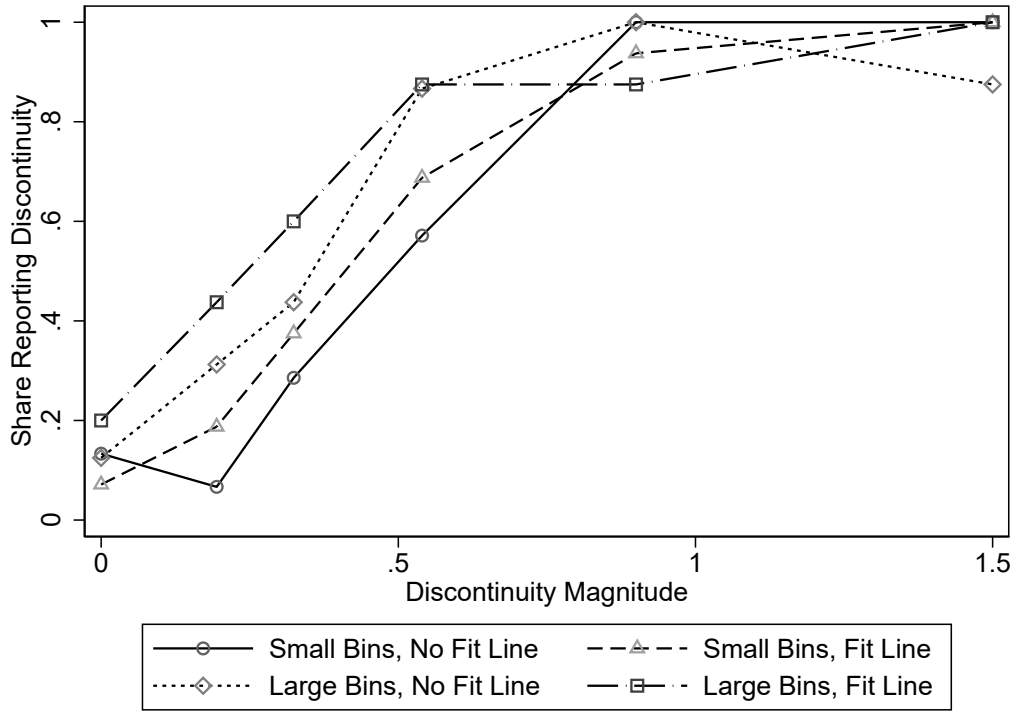
Notes: This figure plots the difference in the RK power functions between experts and non-experts from the right panel of Figure 3.

Figure A.14: RD: Expert Preferences and Beliefs regarding Non-Expert Performance



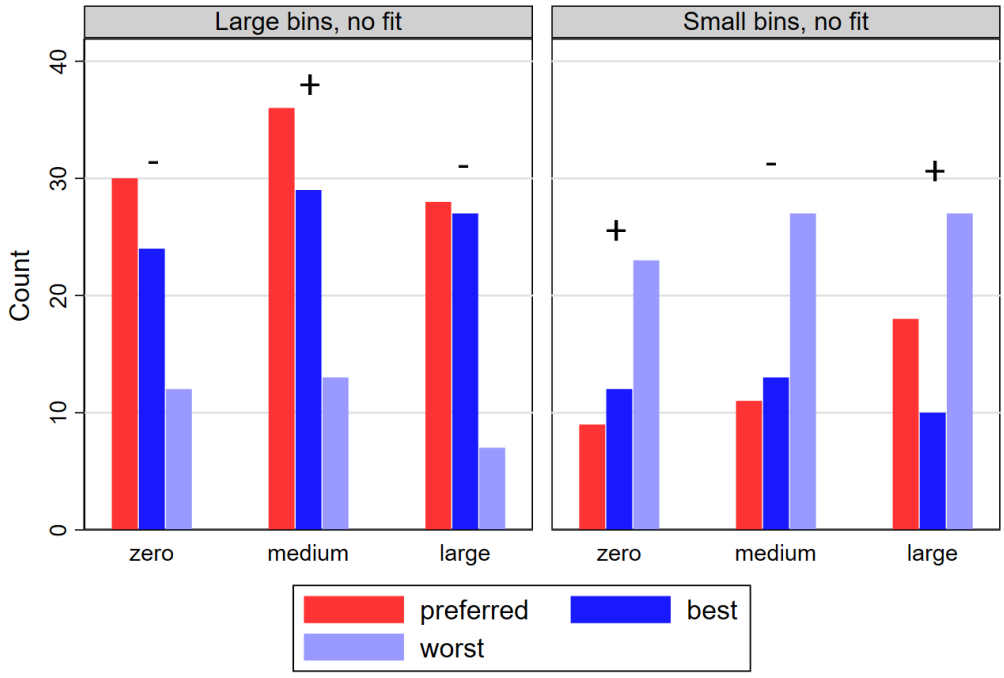
Notes: Each panel of this figure shows the number of experts who report the given treatment as being their preferred treatment at a given discontinuity level; believe it to be the best-performing treatment among our non-expert sample; or believe it to be the worst-performing treatment among our non-expert sample. For comparison, the treatments that performed best and worst in that sample are marked with a $+$ and $-$ sign respectively.

Figure A.15: Phase RDD4: Power Functions for DGP Shown to Experts



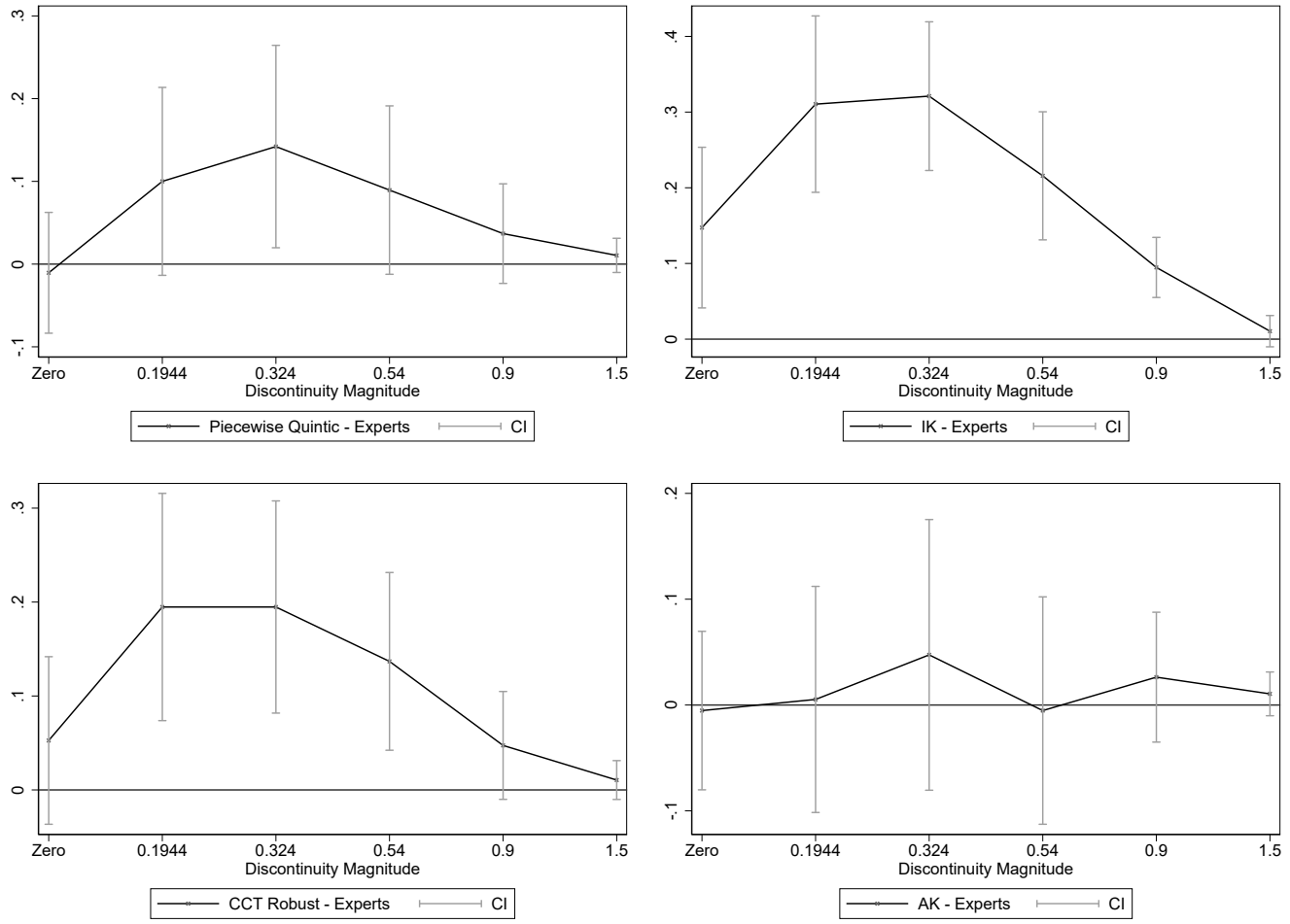
Notes: This figure shows the power functions for the DGP that was used to elicit expert preferences and beliefs across four considered treatments and three discontinuity levels in the second part of the expert survey. The corresponding expert preferences and beliefs are shown in Figure A.14.

Figure A.16: RK: Expert Preferences and Beliefs regarding Non-Expert Performance



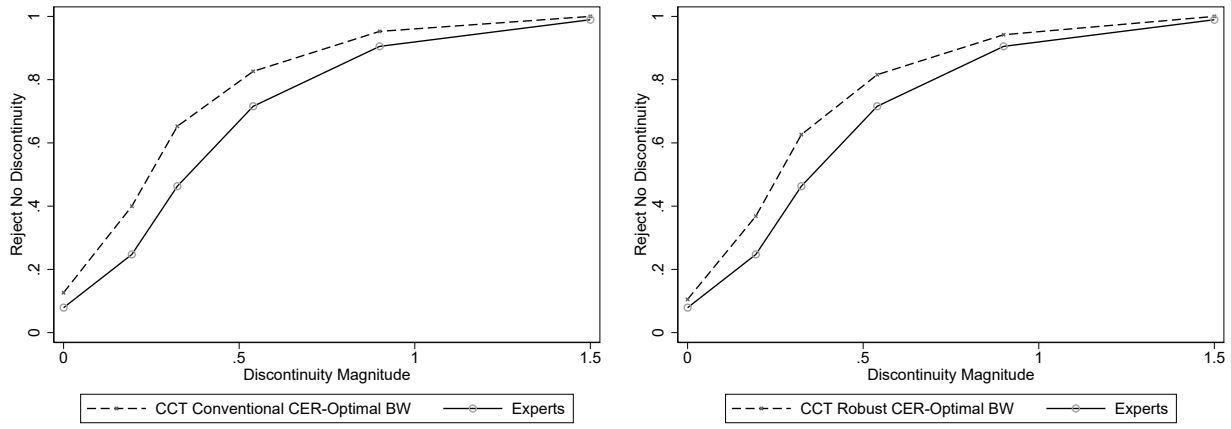
Notes: Each panel of this figure shows the number of experts who report the given treatment as being their preferred treatment at a given discontinuity level; believe it to be the best-performing treatment among our non-expert sample; or believe it to be the worst-performing treatment among our non-expert sample. For comparison, the treatments that performed best and worst in that sample are marked with a $+$ and $-$ sign respectively.

Figure A.17: RD: Expert Visual vs Econometric Inference Power Function Differences



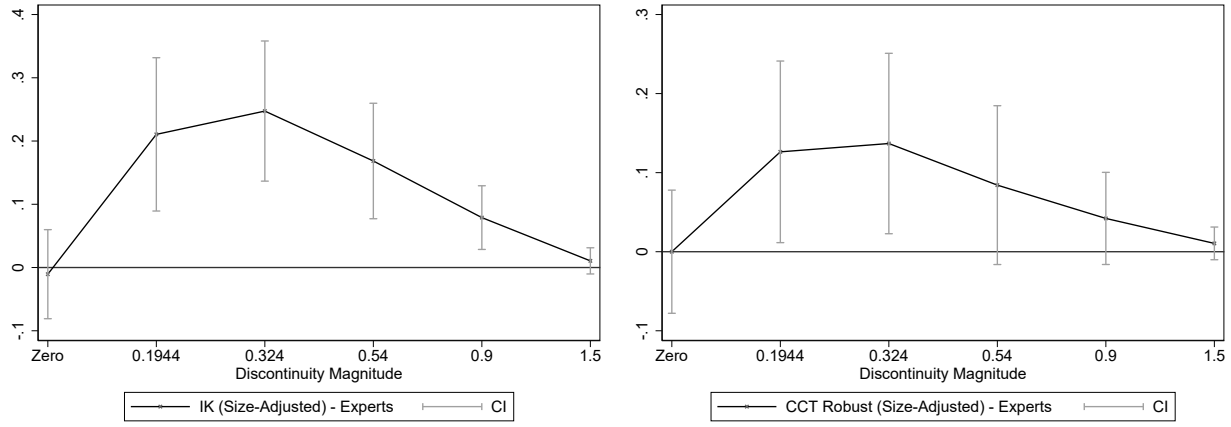
Notes: This figure plots the difference in the RD power functions between experts' visual inference and four econometric inference procedures from the left panel of Figure 4.

Figure A.18: RD: Expert Visual vs CCT Procedure with Inference-Optimal Bandwidth Power Functions



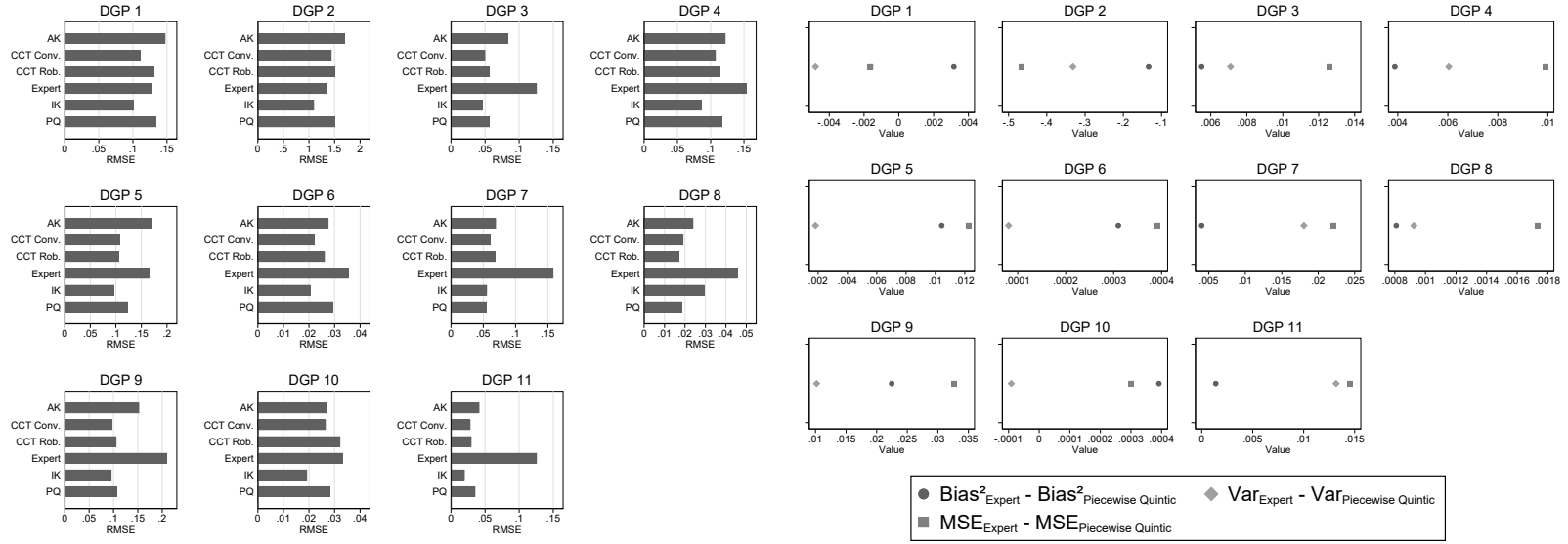
Notes: This figure plots the power functions of experts' visual inference and the CCT procedures with inference-optimal bandwidth per Calonico, Cattaneo, and Farrell (2020)

Figure A.19: RD: Expert Visual vs Size-Adjusted Econometric Inference Power Function Differences



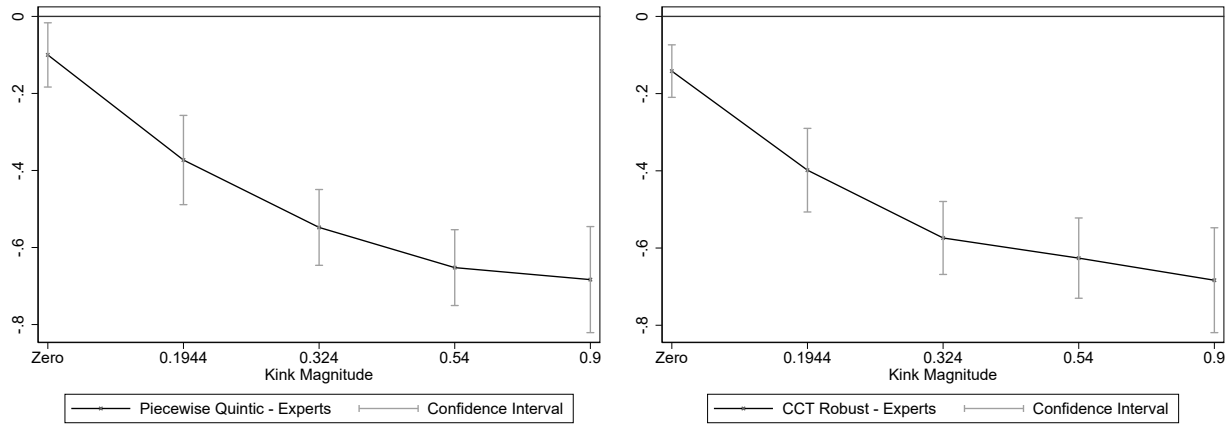
Notes: This figure plots the difference in the RD power functions between experts' visual inference and the size-adjusted IK and CCT procedures from the right panel of Figure 4.

Figure A.20: RMSE of Point Estimates and MSE Decomposition by DGP



Notes: The left panel compares for each DGP the RMSE in estimating the discontinuity magnitude in visual and econometric procedures. The right panel decomposes the difference in RMSE between visual estimation and the correctly specified regression benchmark into the bias and variance components.

Figure A.21: RK: Expert Visual vs Econometric Inferences Power Function Differences



Notes: This figure plots the difference in the RK power functions between experts' visual inference and the default CCT procedure of Figure 5.

Figure A.22: Lineup Protocols for DGPs Specified Before and After Adding Noise to the Running Variable

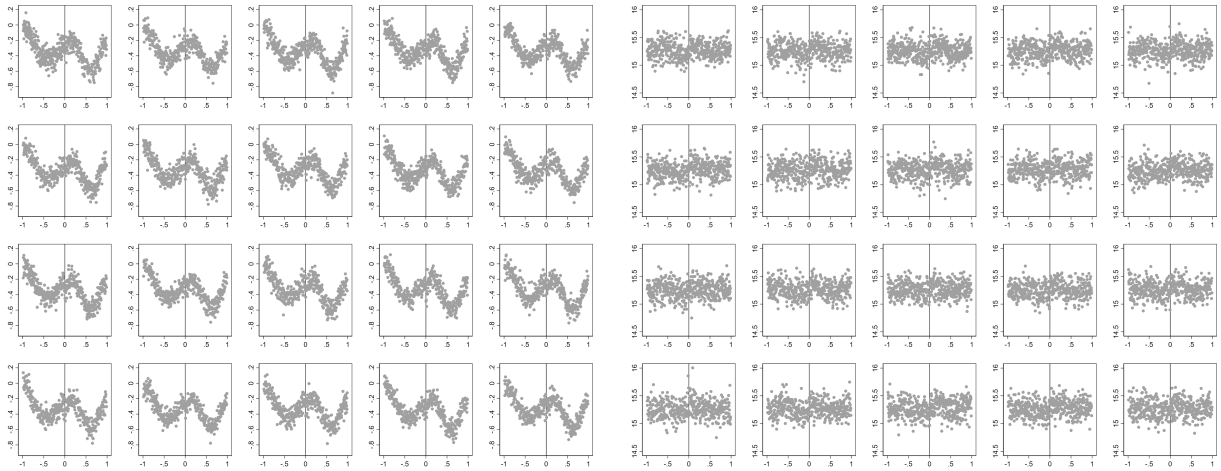


Figure A.23: Comparison between CEFs from Original Microdata and Adding Noise to Running Variable

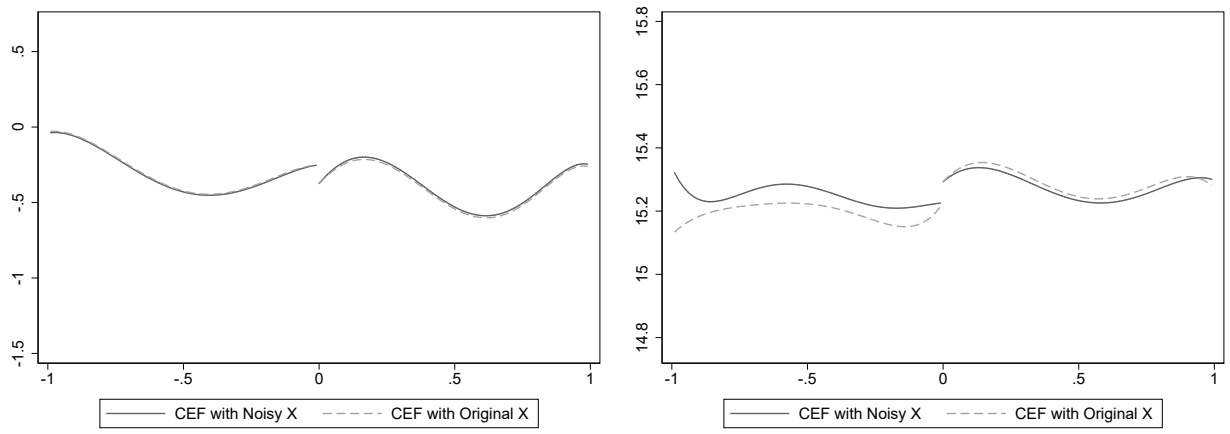


Figure A.24: Video Tutorial Attention Check

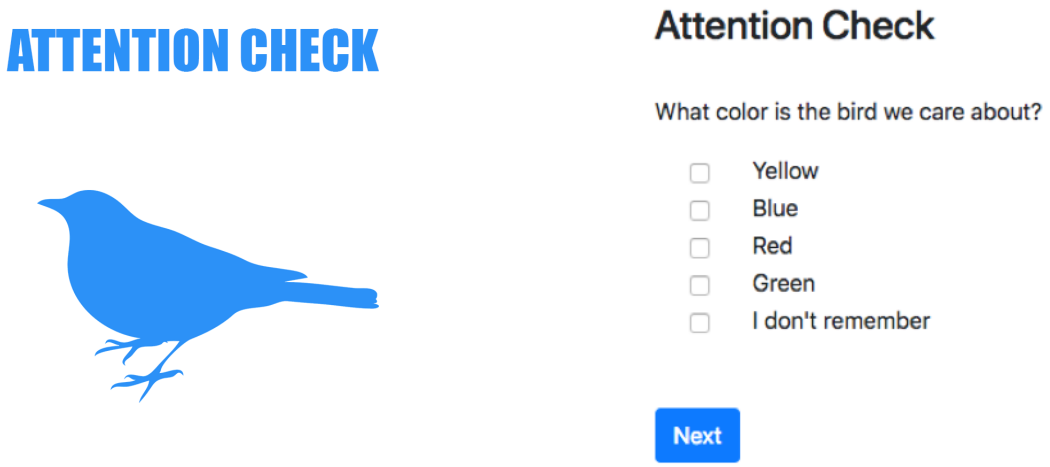
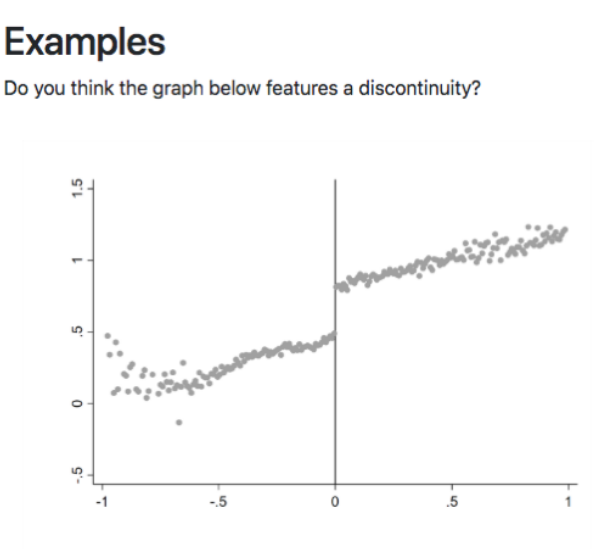


Figure A.25: Example Tasks

(a) Example 1



(b) Example 1 - Feedback

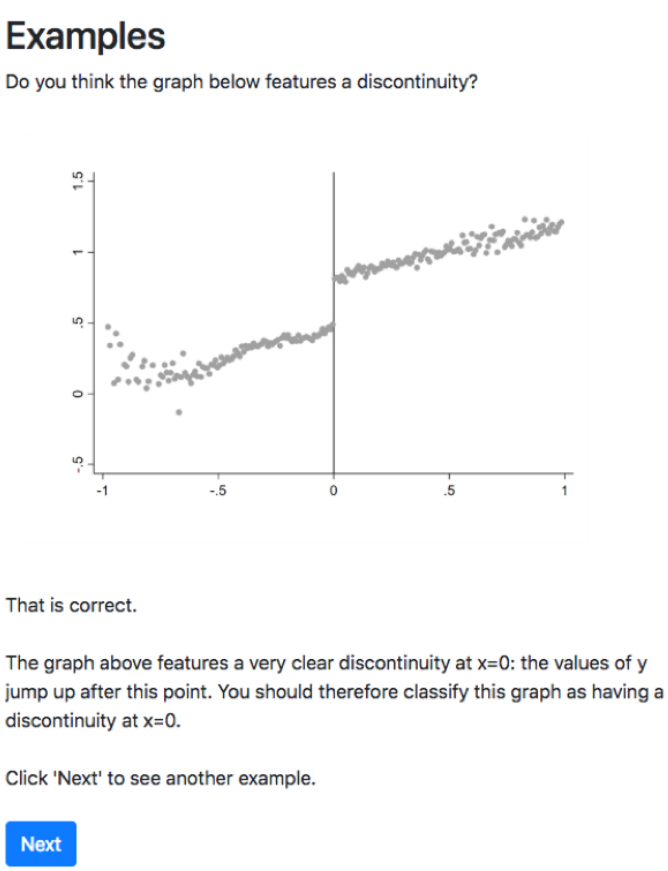
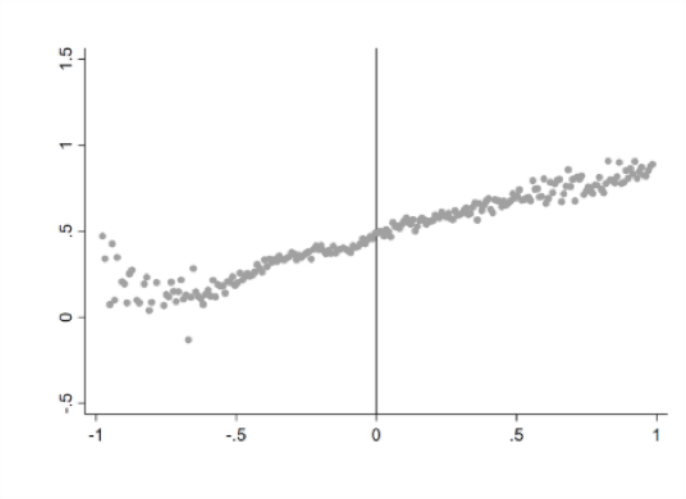


Figure A.26: Example 4 - Feedback and Navigation Buttons

Examples

Do you think the graph below features a discontinuity?



Example 1

Example 2

Example 3

Example 4

This graph is based on the **same** underlying true relationship as the graph in Example 3. However, the data on y are much less variable, making it apparent that **no discontinuity** exists at x=0.

Please click 'Next' when you are ready to view one final example.

Next

Figure A.27: Classification Screen

(a) Incentives

Survey

In this survey, you will be asked to classify 11 graphs depending on whether you believe the underlying true relationship features a discontinuity at x=0 or not.

Please note that all, some or none of the 11 graphs you see in this survey may feature a discontinuity.

For each graph, you have two bonus options. You can select to be paid

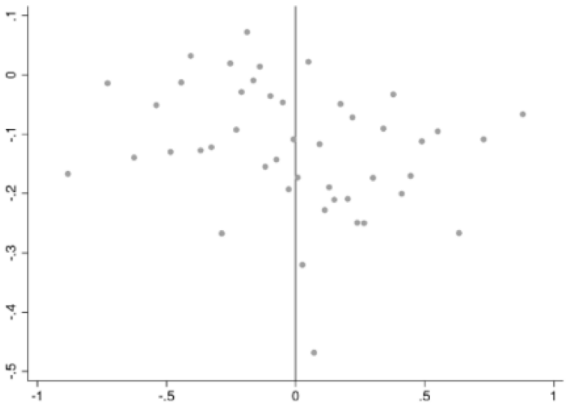
- $0.40 if you answer correctly and $0.00 otherwise, or
- $0.20 if you answer correctly and $0.20 otherwise.

Please click 'Next' to move on to the next page and begin the survey whenever you are ready.

Next

(b) Classification Task

Question 1 of 11



Do you think that there is a discontinuity in the true relationship at x=0?

Yes | No

Please make your bonus selection below:

Payment if Correct	Payment if Incorrect	Selection
$0.40	$0.00	☐
$0.20	$0.20	☐

Figure A.28: Results Screen Example

Results

Thank you for completing this study! You answered 9 questions correctly. Based on your bonus choices, **your total bonus is $3.60**. Your final earnings including the $3.00 participation fee are $6.60.

If you are interested in receiving more information regarding the results of this study, or would like a summary of your answers and earnings, please contact us by email at cmk272@cornell.edu. We will send you the information after the study is completed.

Next

Figure A.29: Sequence of Events - RDD Experiments

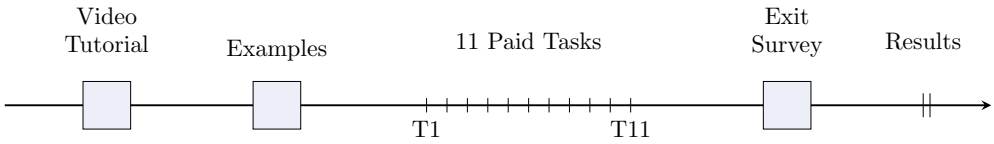
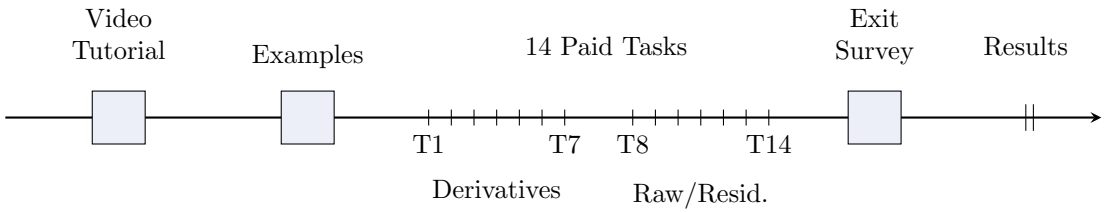


Figure A.30: Sequence of Events - RKD Experiments

(a) Pilot RKD



(b) Phase RKD

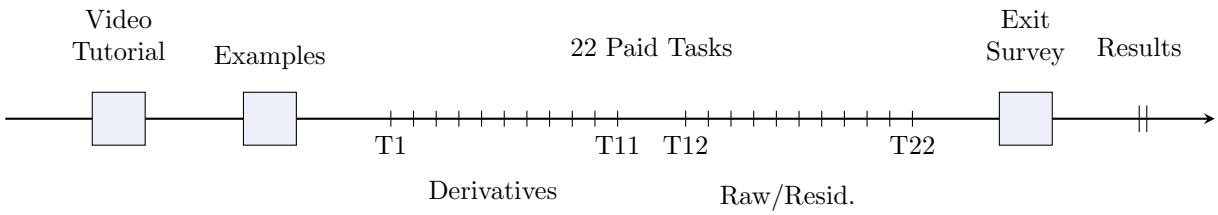


Figure A.31: Sequence of Events - Expert Study

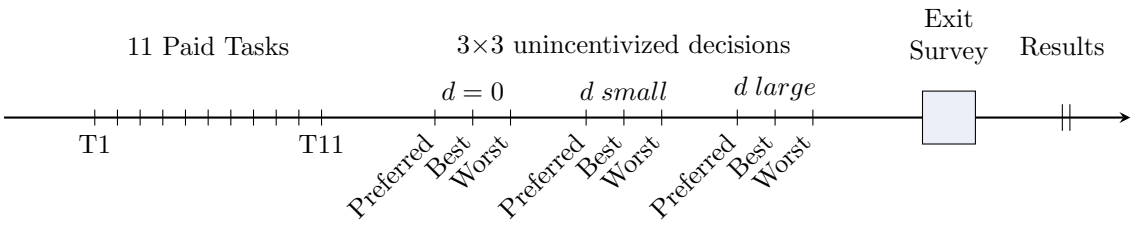


Figure A.32: Expert Survey - Part 1 Introduction

Part 1

You are now ready to begin Part 1 of this survey.

In this part of the survey, we will present you with 11 graphs. You can navigate between graphs using a series of buttons at the bottom of the screen. For each graph, we will ask you to assess whether a discontinuity is present.

If you think that a graph exhibits a discontinuity, we will also ask you to determine the direction of the jump, and to visually estimate the size of the discontinuity with information on the y-axis.

Please note that you will not be able to move on to Part 2 until you have completed Part 1.

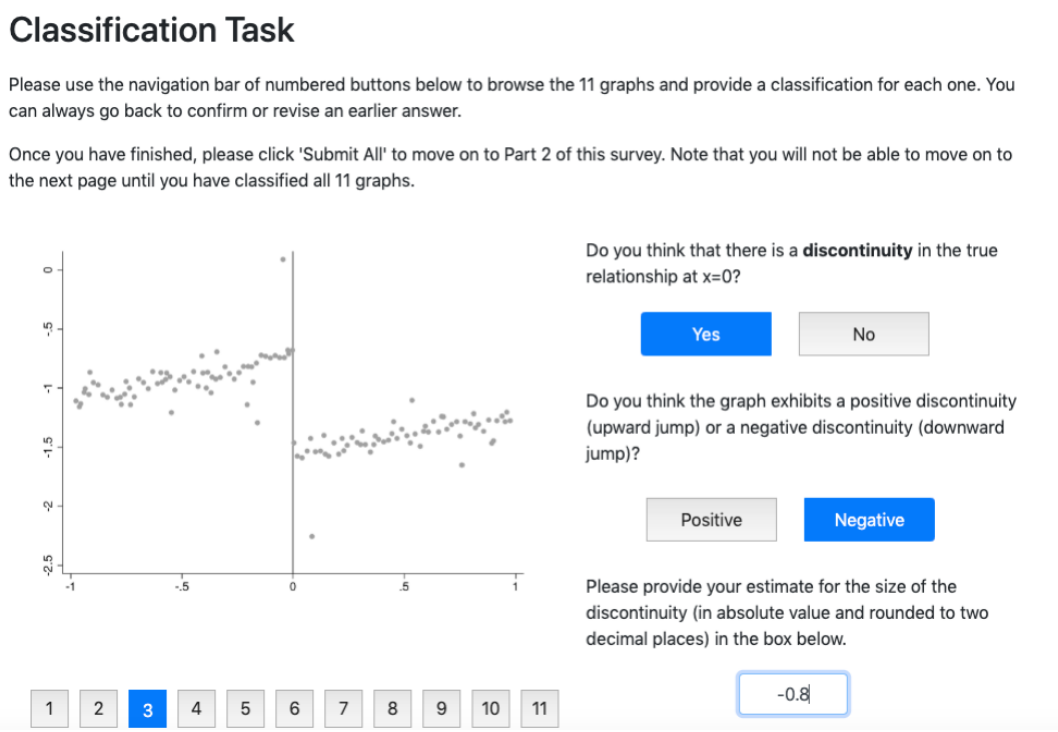
If you are one of the four randomly selected participants at the end of this study, you will earn a base fee of $450 plus $50 for every graph you classify correctly. To earn the bonus, you only need to correctly determine whether or not a discontinuity is present.

Please click 'Next' when you are ready to begin the classification task.

Next

Figure A.33: Expert Survey - Part 1 Classification Task

(a) RD Case



(b) Expert Survey - Part 1 Task Screen (RK case)

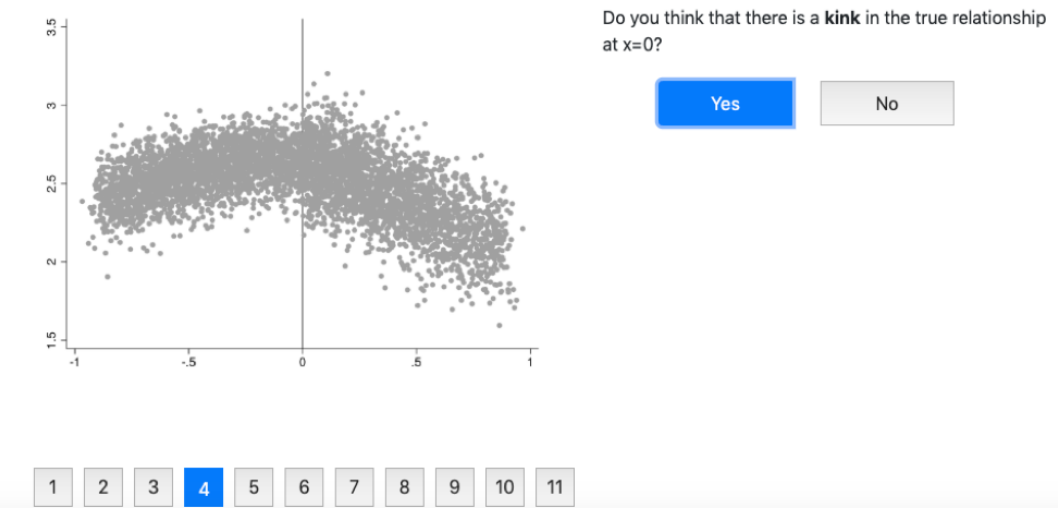
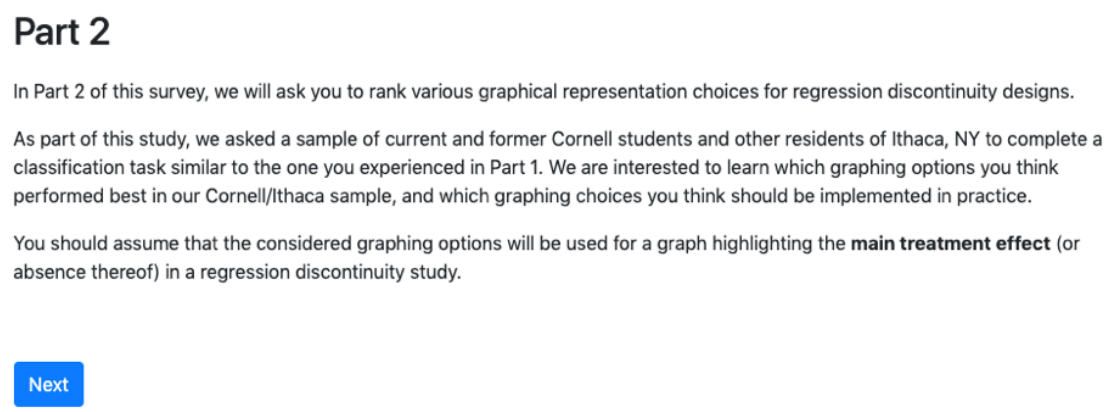


Figure A.34: Expert Survey - Part 2 Instructions

(a) RD Case



(b) Expert Survey - Part 1 Task Screen (RK case)

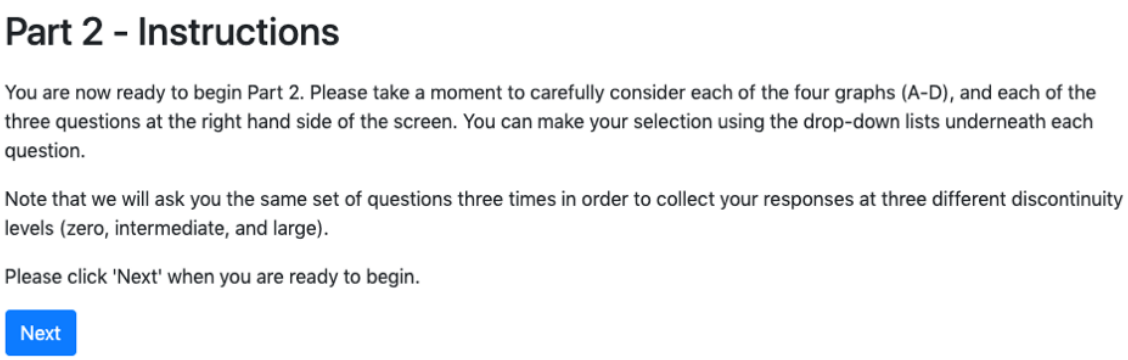


Figure A.35: Expert Survey - Part 2 Decision Screen

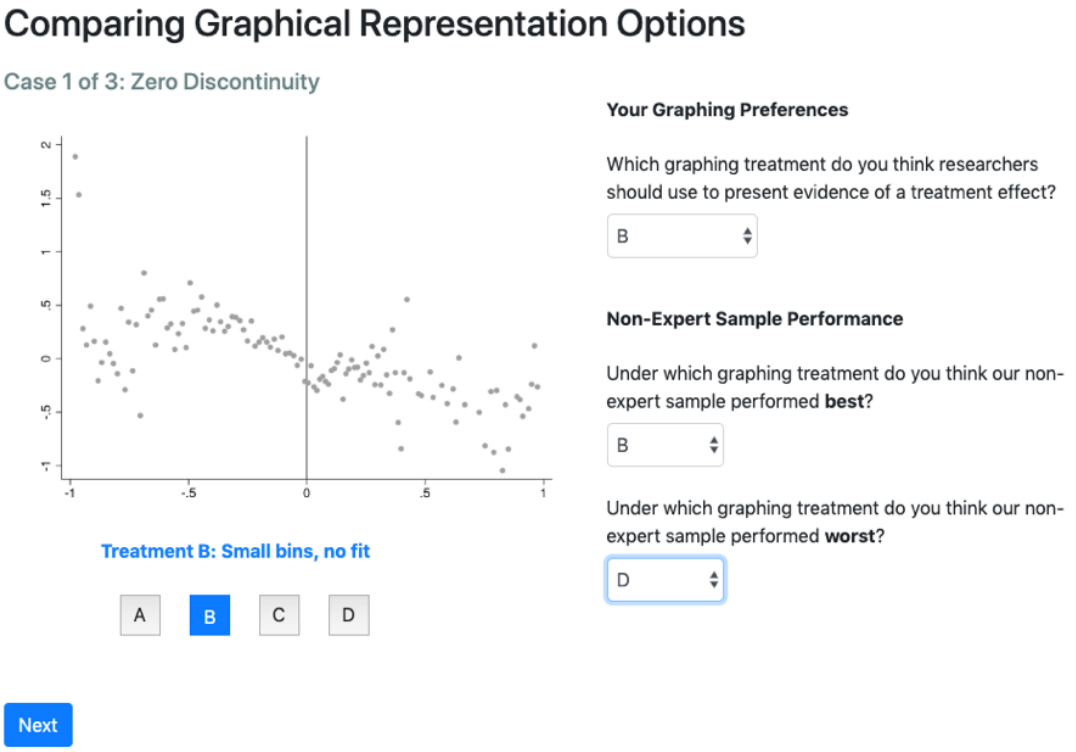


Figure A.36: Sequence of Events - Eyetracking Study

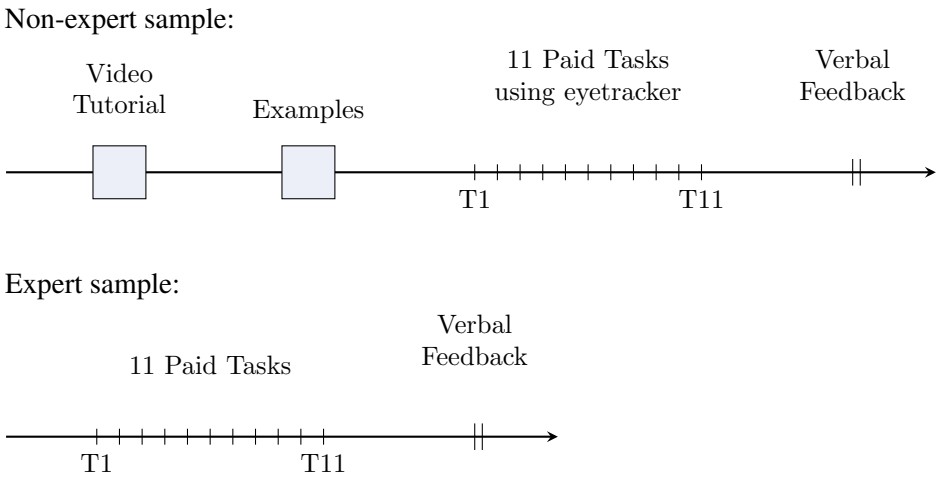


Figure A.37: Sequence of Screens for Each Classification Task on Eyetracking Terminal

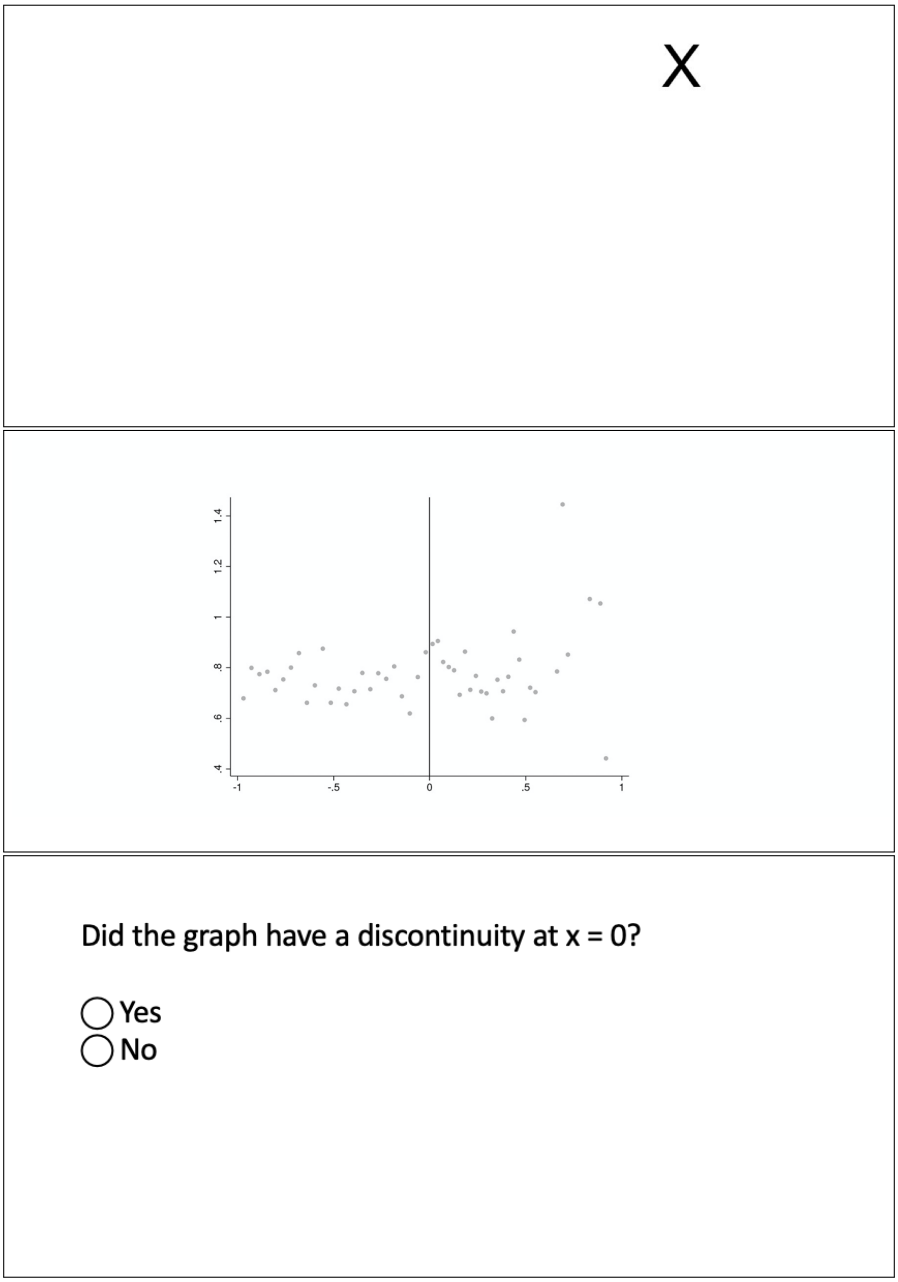
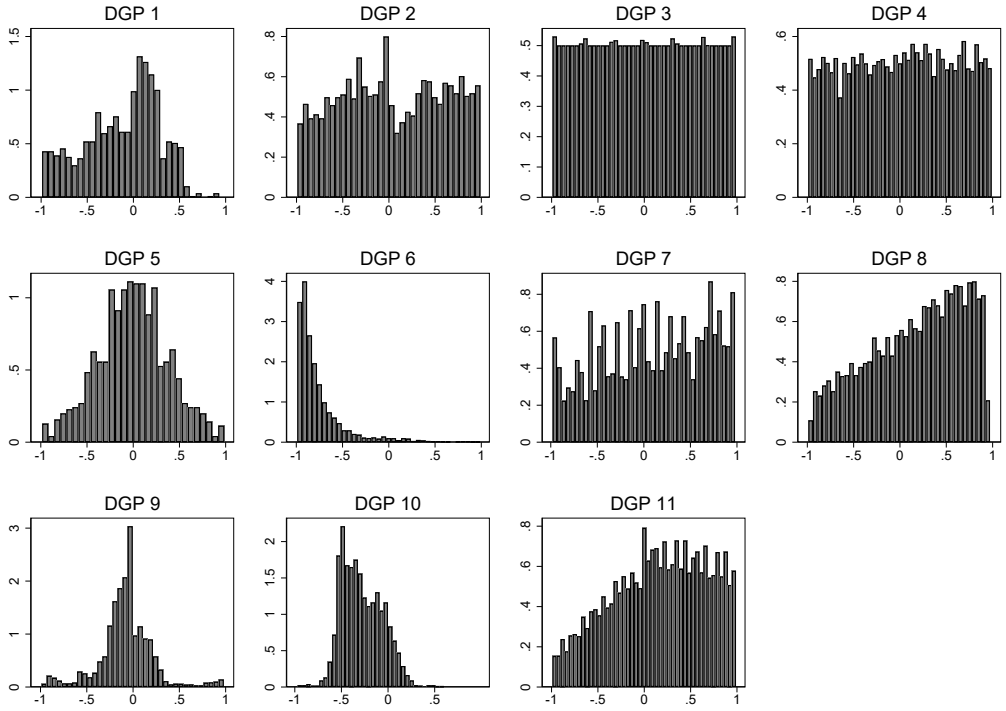
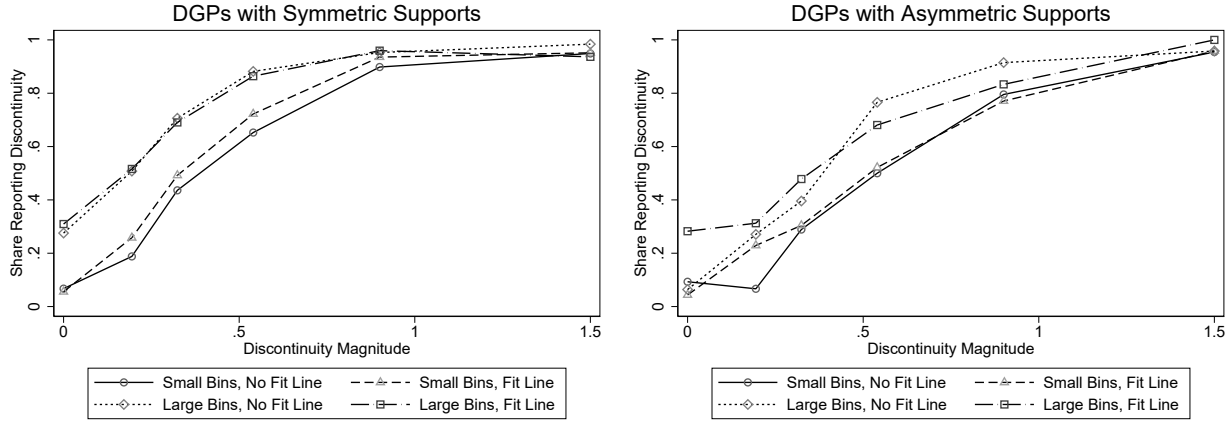


Figure A.38: RDD DGP Histograms



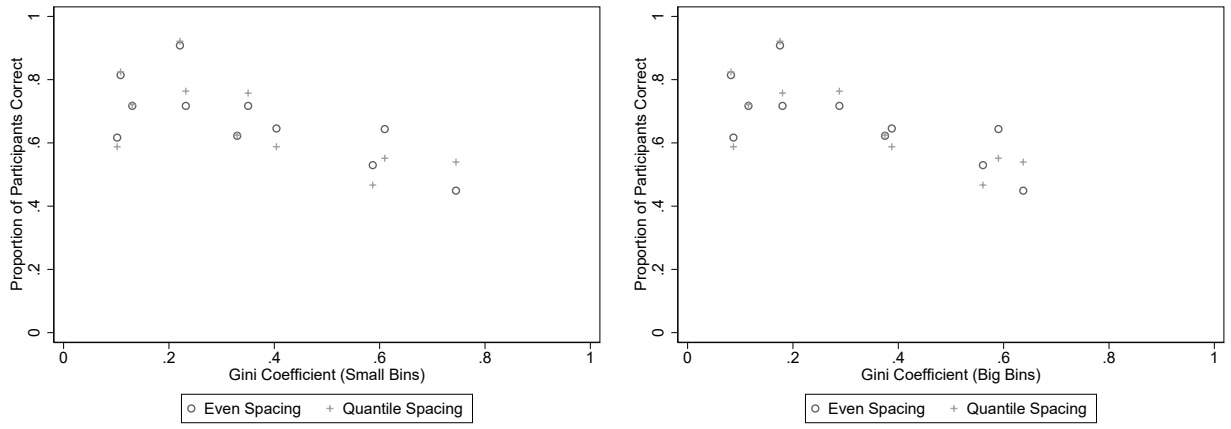
Notes: This figure presents histograms of the running variable for each RDD DGP.

Figure A.39: Comparison between Power Functions for Phase RDD4 by Symmetry of DGP Supports



Notes: This figure plots the power functions from Phase RDD4 broken down into DGPs with symmetric supports prior to normalization and those with asymmetric supports prior to normalization.

Figure A.40: DGP Gini Coefficients and Non-Expert Performance



Notes: This figure compares non-expert accuracy in classifying RDD graphs by DGP based on the uniformity of the distribution of the running variable as measured by the Gini coefficient. The left panel contains results for graphs with small bins, and the right panel contains results for graphs with big bins.

Figure A.41: Discontinuity Easier to Detect When Both Slopes at Cutoff are in the Same Direction as the Discontinuity

(a) Positive Slopes and Positive Discontinuity

(b) Positive Slopes and Negative Discontinuity

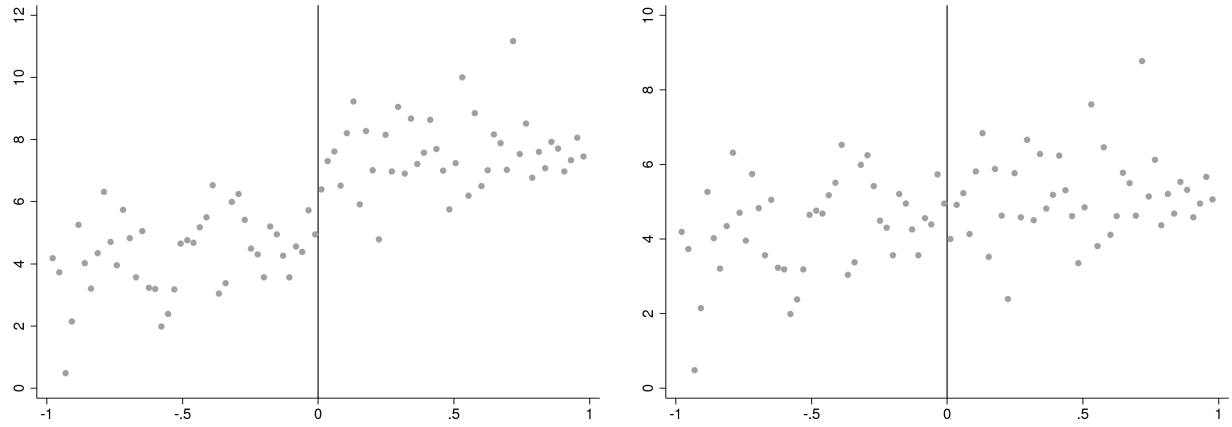
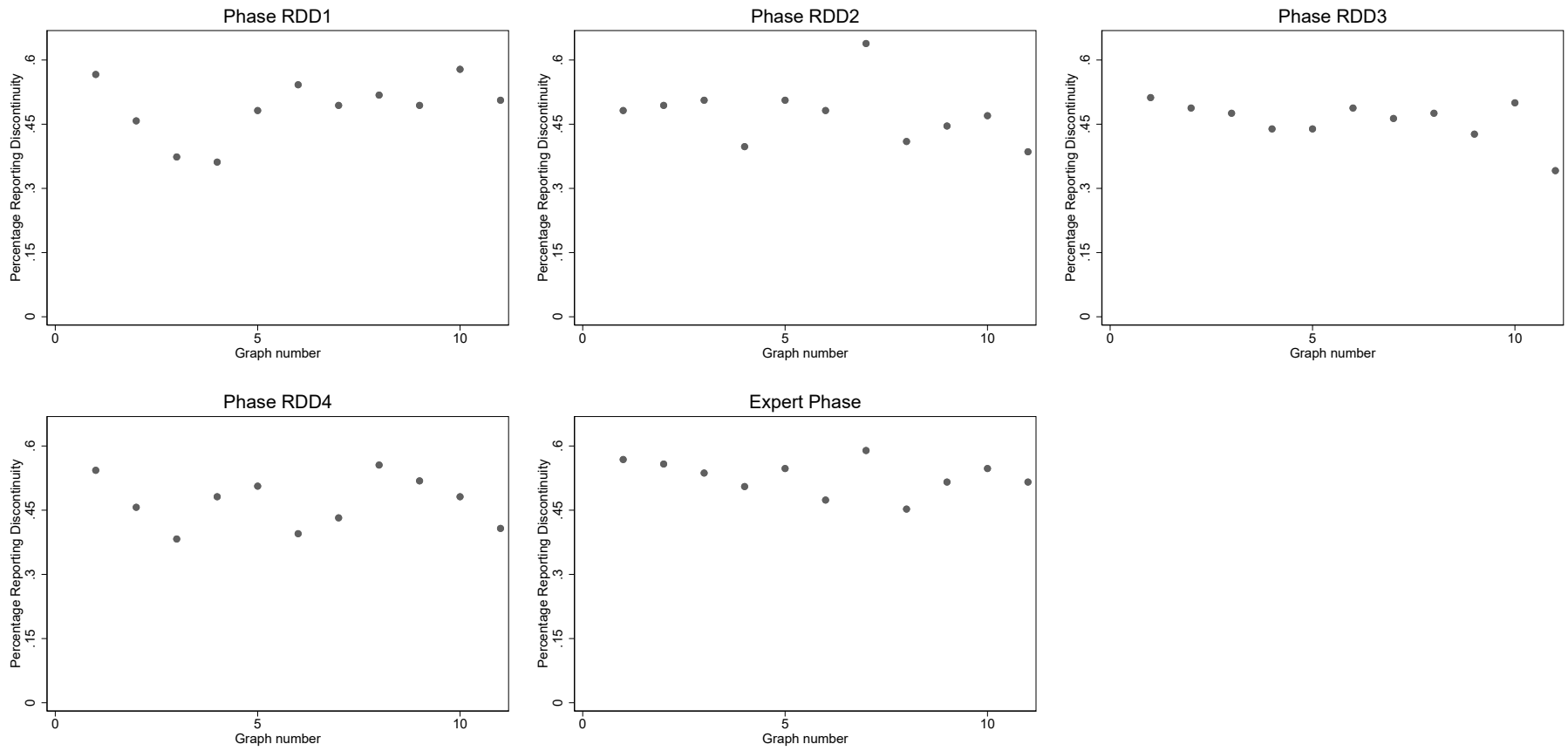
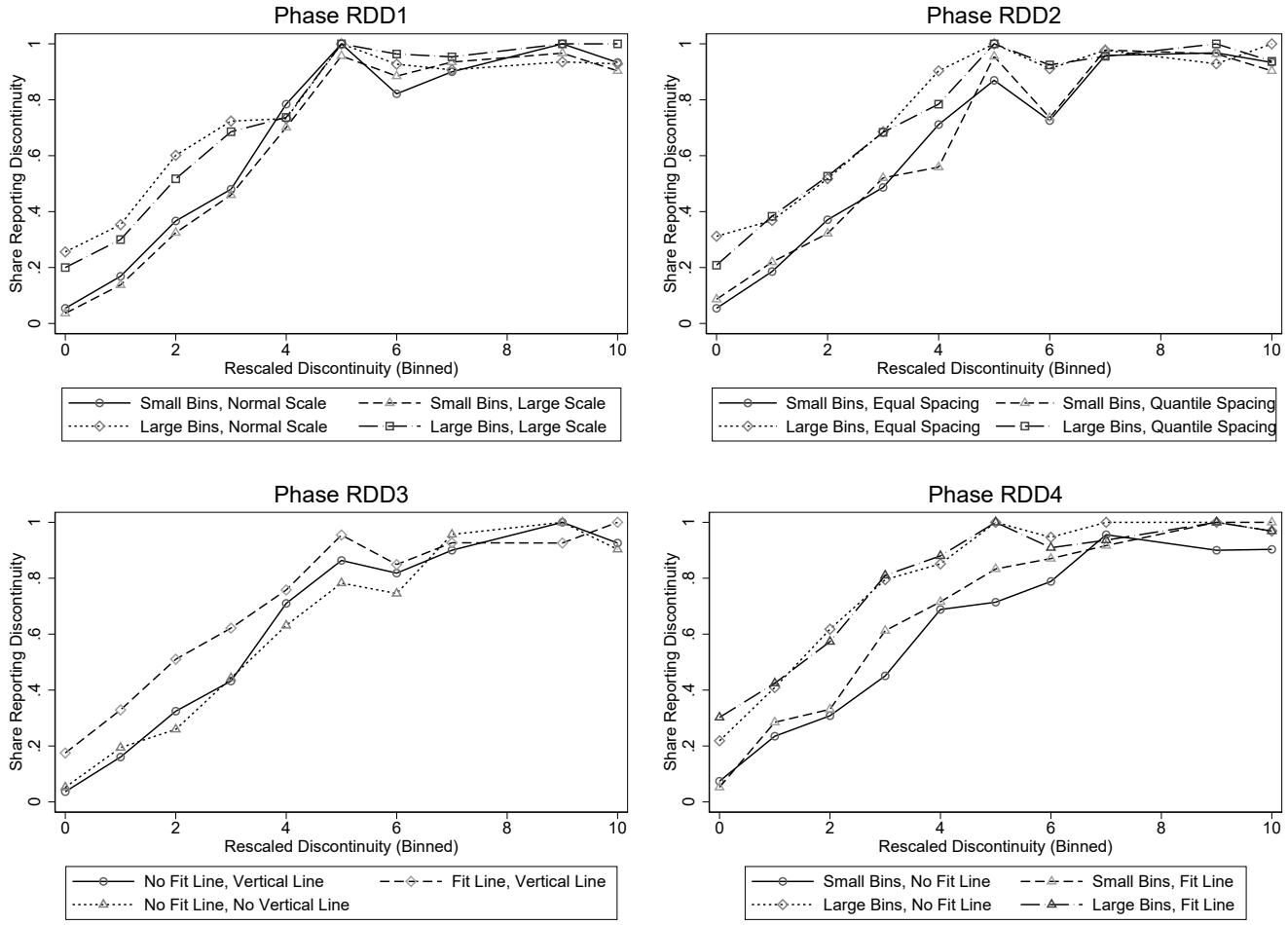


Figure A.42: RDD Discontinuity Classifications over 11 Graphs



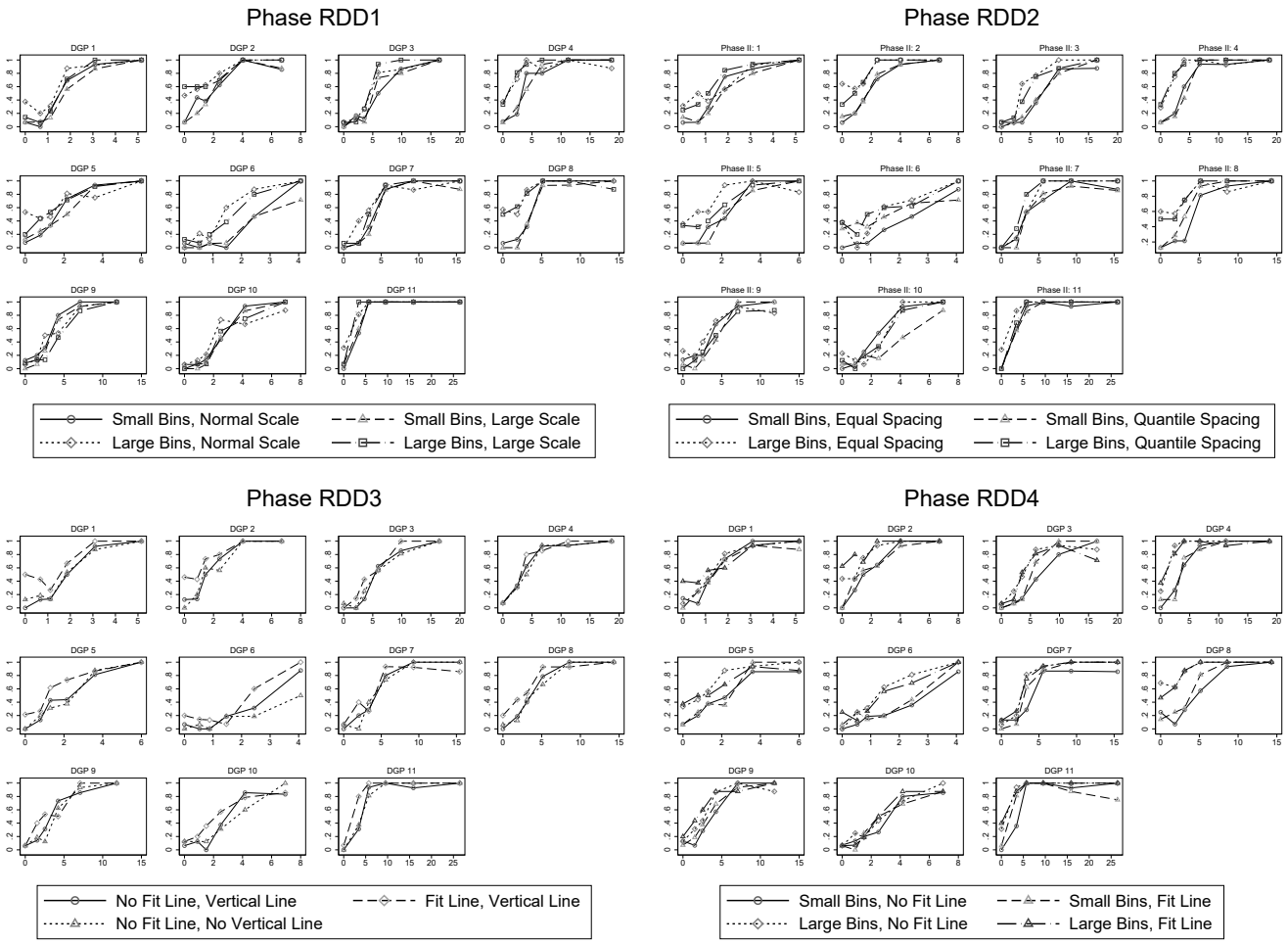
Notes: This figure plots participants' likelihood of reporting a discontinuity over the course of the 11 graphs seen.

Figure A.43: RDD Phases: Power Functions against Asymptotic t-Statistics



Notes: Plotted are empirical power functions from the four non-expert RDD experiments. The power functions are defined in Section 2. The x-axis reflects binned values of the asymptotic t-statistic discussed in Appendix E. The y-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure A.44: RDD Phases: Power Functions by DGP against Asymptotic t-Statistics



Notes: Plotted are empirical power functions from the four non-expert RDD experiments broken down by DGP per phase. The power functions are defined in Section 2. The x-axis represents binned values of the asymptotic t-statistic discussed in Appendix E. The y-axis represents the share of respondents classifying a graph as having a discontinuity at the policy threshold.

Figure A.45: Visual Representation of Eyetracking Data

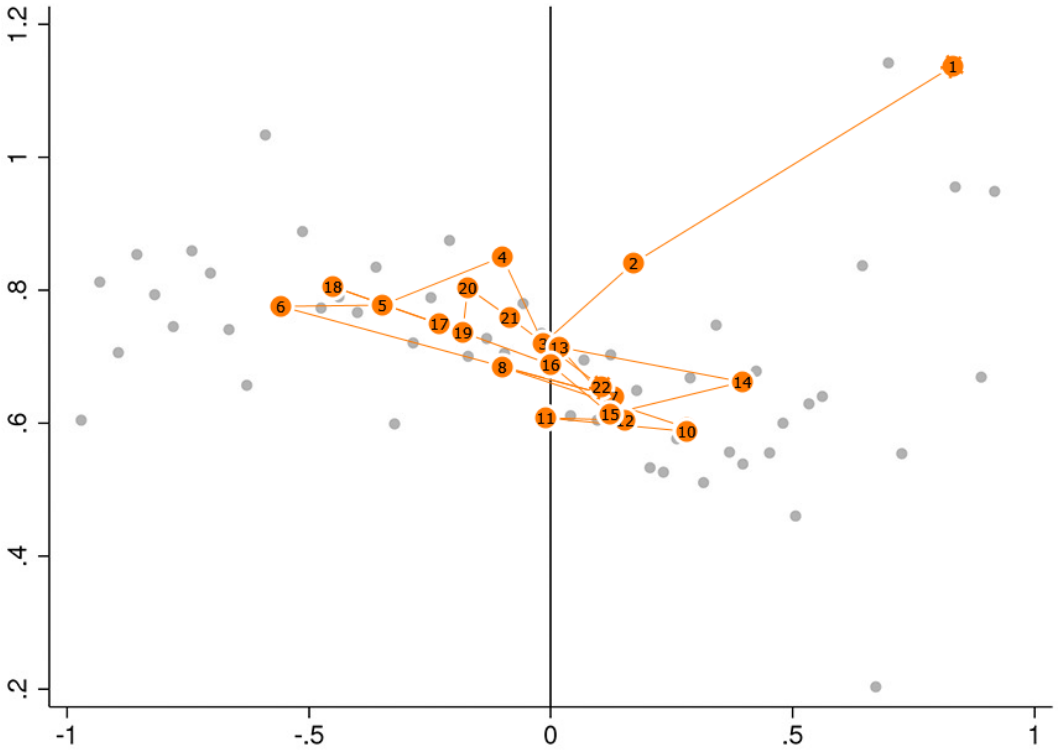
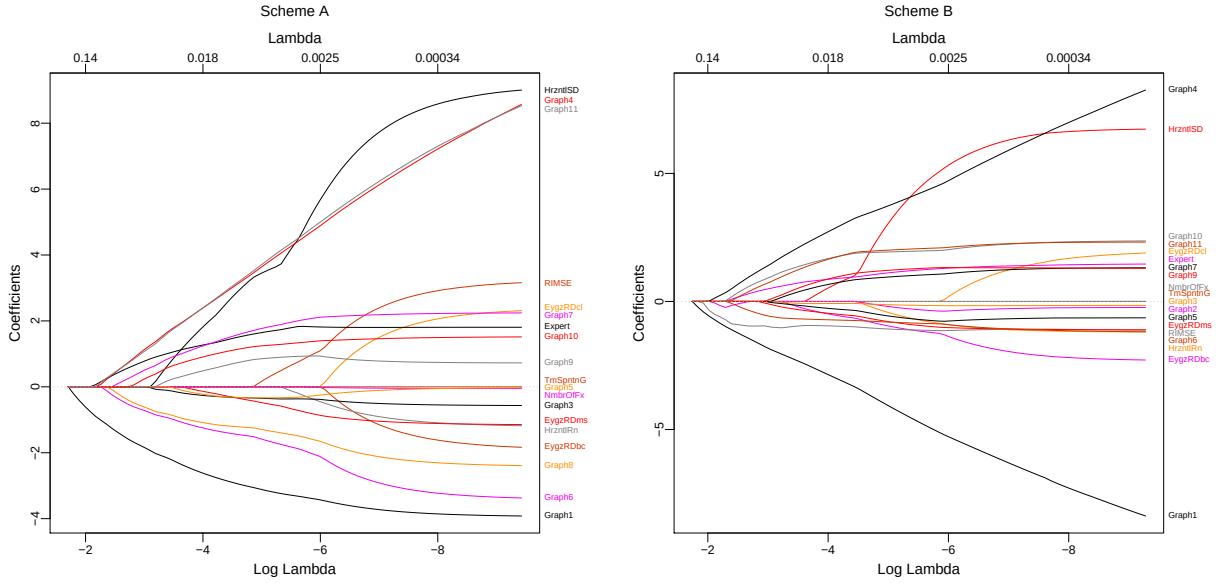
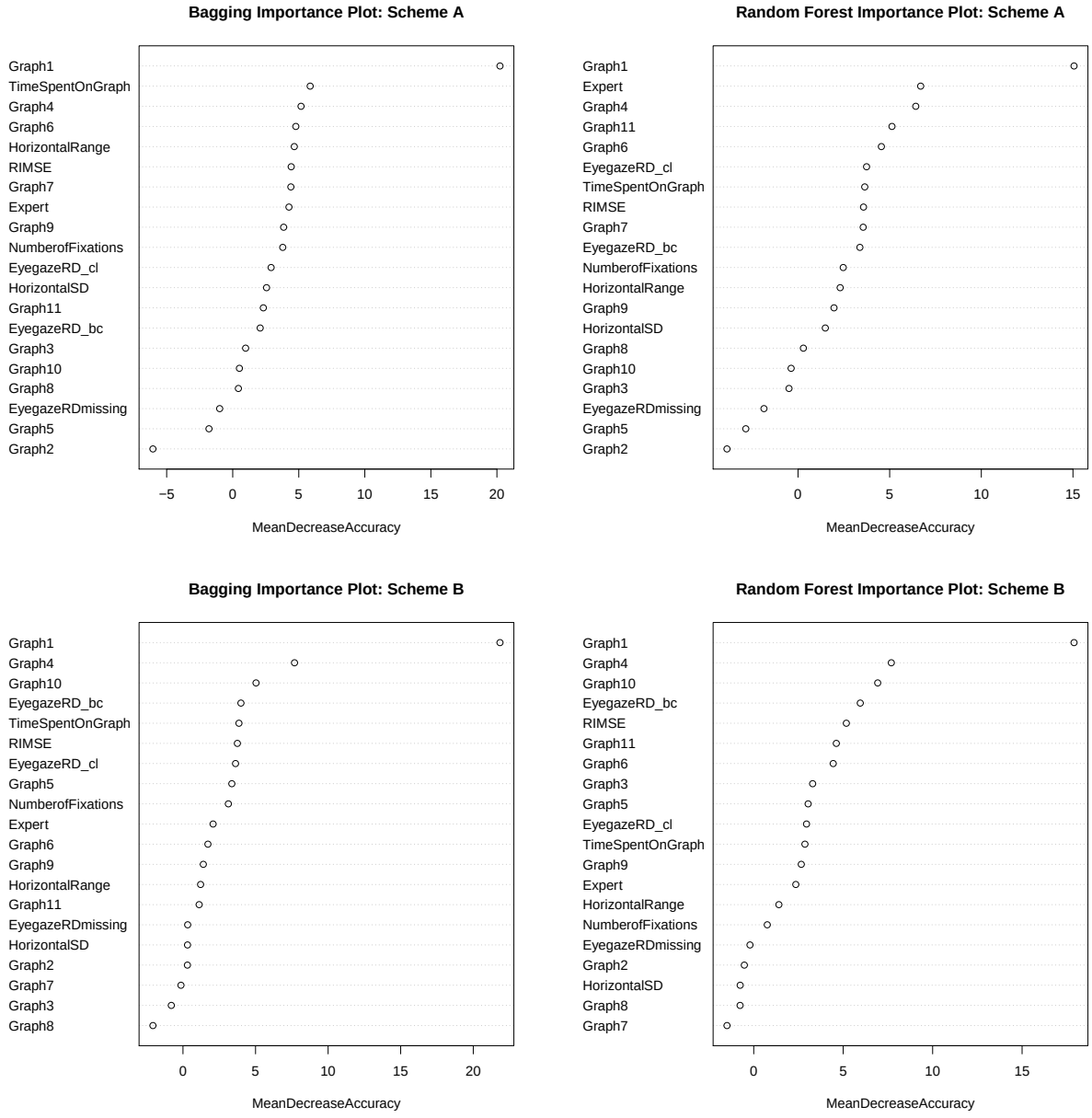


Figure A.46: Penalized Logit Coefficient Path Plots



Notes: For the abbreviated variables: "TmSpntnG" reflects the time spent per graph in seconds; "NmbrOfFx" is the number of fixations per graph; "HrzntlRn" is the horizontal range of eyegazes; "HrzntlSD" is the horizontal standard deviation of eyegazes; "EygzRDcl" is the absolute deviation of the conventional local linear RD estimates from the true discontinuity (scaled by σ); "EygzRDbc" is the absolute deviation of bias-corrected local linear RD estimates from the true discontinuity; and "EygzRDms" reflects whether the discontinuity estimate is unavailable.

Figure A.47: Bagging and Random Forest Importance Plots



Notes: For the abbreviated variables: "EyegazeRD_cl" reflects the absolute deviation of conventional local linear RD estimates from the true discontinuity (scaled by σ); "EyegazeRD_bc" is the absolute deviation of bias-corrected local linear RD estimates from the true discontinuity; and "EyegazeRDmissing" reflects whether the discontinuity estimate is unavailable.