

VISUAL INFERENCE AND GRAPHICAL REPRESENTATION IN REGRESSION DISCONTINUITY DESIGNS*

Christina Korting
Carl Lieberman
Jordan Matsudaira
Zhuan Pei[†]
Yi Shen

Total Word Count: 10,367

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 study visual inference within the context of regression discontinuity (RD) designs by measuring how accurately readers identify discontinuities in graphs produced from data-generating processes calibrated on 11 published papers from leading economics journals. First, we assess the effects of different graphical representation methods on visual inference using randomized experiments. We find that bin widths and fit lines have the largest impacts on whether participants correctly perceive the presence or absence of a discontinuity. Our experimental results allow us to make evidence-based recommendations to practitioners, and we suggest using small bins with no fit lines as a starting point to construct RD graphs. Second, we compare visual inference on graphs constructed using our preferred method with widely used econometric inference procedures. We find that visual inference achieves similar or lower type I error (false positive) rates and complements econometric inference.

Key Words: Graphical Methods; Visual Inference; Regression Discontinuity Design; Expert Prediction; Scientific Communication

JEL Code: A11, C10, C40

*We are grateful for the insightful and constructive comments from two editors, Larry Katz and Andrei Shleifer, and four anonymous reviewers. We have also benefited from discussions with Alberto Abadie, Sahara Byrne, Colin Camerer, Matias Cattaneo, Damon Clark, Geoff Fisher, Paul Goldsmith-Pinkham, Nathan Grawe, Jessica Hullman, David Lee, Lars Lefgren, Thomas Lemieux, Pauline Leung, Jia Li, Adam Loy, Alex Mas, Doug Miller, Ted O'Donoghue, Bitsy Perlman, Steve Pischke, Jonathan Roth, Jesse Rothstein, Rocio Titiunik, Cindy Xiong, Stephanie Wang, Andrea Weber, and Xiaoyang Ye, as well as participants of various seminars and conferences. Lexin Cai, Matt Comey, Michael Daly, Rebecca Jackson, Motasem Kalaji, Xingyue Li, Fiona Qiu, and Tatiana Velasco provided research assistance, and we thank Brad Turner, Mary Ross, and Patti Tracey for providing logistical support. We are indebted to our friends for testing the experiment and to colleagues for participating in the study. We gratefully acknowledge financial support from the Cornell Institute of Social Sciences and the Princeton Industrial Relations Section. 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. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau.

[†]Corresponding author. Associate Professor, Department of Economics and Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY 14853, USA; zhuan.pei@cornell.edu

1 Introduction

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

— Geoffrey S. Watson (1964)

Graphical analysis is increasingly prevalent in empirical research, a phenomenon Currie, Kleven, and Zwiers (2020) call the “graphical revolution.” Effective use of graphs conveys a large set of statistical information at once and improves research transparency (Andrews, Gentzkow, and Shapiro, 2020). However, there are different ways to construct a graph with the same data, and the particular construction an analyst chooses has the potential to mislead readers (Schwartzstein and Sunderam, 2021). To understand the best use of graphical evidence, it is important to study readers’ ability to process information from graphs—which we term “visual statistical inference” or *visual inference* per Majumder, Hofmann, and Cook (2013)—as well as the sensitivity of visual inference to choices in graph construction. To date, little is known about visual inference for commonly presented graphs in empirical research designs.

We begin to fill this knowledge gap and study 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 Pischke (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), Lee and Lemieux (2010), and Cattaneo, Idrobo, and Titiunik (2019) 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 accurately extract the embedded statistical information, where our main criterion is the correct identification of the existence or absence of a discontinuity at the policy threshold. Our project has two major components. In the first, we build on work pioneered by Eells (1926) and refined by Cleveland and McGill (1984) by conducting a series of randomized experiments to examine how different graphical parameters affect visual inference in RD. We present participants recruited through the Cornell University Johnson College’s Business Simulation Lab with RD graphs produced from data-generating processes (DGPs) based on

microdata from 11 published papers that we randomly select from a list of 110 empirical studies from top economics journals. 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 popular data-driven 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 advantages and disadvantages of different graphical parameters. We find that certain graphical parameters such as bin width and smoothed regression lines create important tradeoffs between type I errors (identifying a discontinuity when there is none, i.e., a false positive) and type II errors (identifying the absence of a discontinuity when there is one, i.e., a false negative). Relative to MV (small) bins, using IMSE (large) bins tends to increase type I error rates but decrease type II error rates. Similarly, imposing fit lines may also increase type I error rates, echoing the concerns of Cattaneo and Titiunik (2021a,b).

To translate our findings to recommendations on graphical practices in RDDs, we empirically implement a decision-theoretic framework that incorporates classification accuracy as a metric to compare graphical methods. The method that uses MV bins with no fit lines consistently performs well relative to IMSE bins and/or imposing fit lines. 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. Our recommendation is robust to alternative decision-theoretic framework formulations.

Because only nonexperts participate in our randomized experiments on the effects of graphical methods, one may be concerned that our results are less relevant to academic audiences. To assess whether our findings generalize across experience levels, we also recruit experts from a pool of seminar attendees and affiliates of the National Bureau of Economic Research (NBER) and the Institute of Labor Economics (IZA)

to participate in our study. Although our expert sample is not large enough to conduct the same randomized experiments, we can compare the nonexpert and expert results by using the subset of nonexperts who saw the same graphs as the experts. We find that the two groups perform comparably.

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 nonexperts. We find that experts only partly anticipate the aforementioned effects of bin widths and fit lines.

As a second major component of the project, we compare the performances of visual inference and econometric inference. For visual inference, we use results from the sample of experts who viewed graphs constructed with the best-performing technique from our experiments. For econometric inference, we apply three influential methods by Imbens and Kalyanaraman (2012), Calonico, Cattaneo, and Titiunik (2014), and Armstrong and Kolesár (2018) (henceforth IK, CCT, and AK, respectively) and conduct hypothesis testing at the 5% (asymptotic) level. We find that visual inference achieves a type I error rate that, at just below 8%, is lower than the IK and CCT procedures (the CCT type I error rate is not significantly higher), but the two econometric procedures enjoy considerably lower type II error rates. Visual inference performs very similarly to the AK procedure, a remarkable result given the minimax optimality property of AK.

We also find that visual and econometric inferences appear to be complementary. First, we examine the joint distribution of visual and econometric tests: while they commit similar type II errors, there does not appear to be a strong association in their type I errors. Second, we assess the performance of a combined visual and econometric inference. One simple way of combining the two inferences mirrors an approach in which a researcher believes a discontinuity exists if and only if a formal test rejects the null hypothesis of no effect *and* she sees a discontinuity in the RD graph. We find that the combined IK and visual inference performs similarly to the AK procedure, which may help explain the enduring credibility of the RD design despite formal inference issues in earlier RD papers.

Finally, we ask experts to estimate the discontinuity magnitude when they classify a discontinuity, and we compare the accuracy of their estimates to that of econometric methods. On this front, econometric methods tend to do better. For example, the simple local linear IK estimator yields lower mean squared errors than experts across all 11 DGPs, shedding light on the limits of visual inference.

This paper connects a diverse set of literatures and makes the following contributions. First, we begin to fill an important gap in our understanding of graphical evidence by evaluating visual inference and graphical representation practices in a widely used quasi-experimental research design. Our endeavor draws

from three strands of the statistics literature that study the choice of graphical parameters (e.g., Calonico, Cattaneo, and Titiunik, 2015, Li et al., 2020), their effects on visual inference (e.g., Cleveland and McGill, 1984), and the evaluation of visual inference through comparison with econometric inference (e.g., Majumder, Hofmann, and Cook, 2013). Our paradigm can be applied to other important areas, as discussed in Section 5.

Second, to guide our study design and to help interpret our empirical results, we propose a general conceptual framework, which may extend to future studies of visual inference in other contexts. In particular, we can interpret the average type I or II error rate we use as an estimate of the probability that a randomly sampled reader commits such an error when viewing a graph generated from a randomly chosen DGP. We show that these error probabilities are key inputs in a standard decision-theoretic framework, which helps inform best graphical practices. We also discuss alternative decision-theoretic formulations and make connections to the recent literature on scientific communication (e.g., Andrews and Shapiro 2021).

Third, 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.

We introduce the conceptual framework in Section 2, describe the design of our experiments and studies in Section 3, present results in Section 4, and conclude in Section 5. For readers in a hurry, the key takeaway results are in Figures V and VIII with corresponding discussions in Sections 4.1 and 4.3.

2 Conceptual Framework

In this section, we propose a conceptual framework for evaluating visual inference to guide our study design and aid in the interpretation of our empirical results. More specifically, we show how to aggregate visual inference performances across subjects, who may reach different conclusions even when viewing the same graph, and how to meaningfully interpret the parameter to which our aggregate measure corresponds. This performance measure helps to inform best graphical practices.

Although RD graphs may serve other purposes, we view their most important function as accurately conveying discontinuity existence and 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 also motivated 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 focus on binary classifications of a graph and treat type I and type II errors as the main performance measures for visual inference. The conceptual framework easily generalizes to assessing visual estimates of the discontinuity magnitude, which we elicit from experts. 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.

To define our aggregate measures of visual inference performance, we introduce the following notations. First, the vector γ denotes a combination of graphical parameters (see Wilkinson, 2013 for an extensive list). We study five parameters in this paper (bin width, bin spacing, axis scaling, polynomial fit lines, and a vertical line at the policy threshold), and each of the five entries of γ represents the value of a particular parameter.¹

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 the policy threshold $x = 0$; iii) the distribution of the error term u where $\tilde{Y} \equiv E[\tilde{Y}|X = x] + u$; and iv) the sample size N . 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, $Y = \tilde{Y} + d \cdot 1_{[X \geq 0]}$ (which implies that $E[Y|X = x] = E[\tilde{Y}|X = x] + d \cdot 1_{[X \geq 0]}$; we provide a graphical illustration in Section 3.2). 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. Typically, the (g, d) combination is jointly referred to as the “data-generating process,” but we separate the discontinuity level d and call g the DGP for ease of exposition below.

We think of each (bivariate) RD dataset as a realization from the probability model (g, d) , which we denote by W , or $W(g, d)$ if we want to emphasize the underlying probability model. Implementing a graphical procedure with parameters γ on dataset W results in an RD graph (γ, W) , which we denote by T or $T(\gamma, g, d)$.

¹In 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.

Alternatively, we can think of T as a realization from (γ, g, d) and refer to (γ, g, d) as the graph-generating process (GGP).

When presented with the same RD graph, readers may draw different visual inferences. For example, some readers 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. We use ϕ to capture these human characteristics that affect graph perception.

The probability that a reader with characteristics ϕ reports that a discontinuity exists in RD graph $T(\gamma, g, d)$ is denoted by $\tilde{p}(T(\gamma, g, d), \phi)$. From casual observation, we know that the same reader 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 we use to gauge reader performance. First, averaging \tilde{p} over both data realizations W and reader characteristics ϕ leads to the quantity

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

This is the probability that a randomly chosen reader reports a discontinuity in a graph randomly generated from the GGP (γ, g, d) . A high value of p indicates a high classification error probability 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 parameters 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, \mathcal{G} (we discuss \mathcal{G} after Assumption 1 below) to arrive at the *overall* discontinuity classification probability for γ :

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

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

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

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

Consistent with the definitions in Casella and Berger (2002, p. 382), we call $p(\gamma, g, d)$ and $\bar{p}(\gamma, d)$ *power functions* as functions of d .

In this paper, we design experiments to estimate the type I and type II error probabilities as defined above. For each GGP (γ, g, d) , we generate M different realized graphs and present each to a random participant. That is, participant i is shown one RD graph denoted by $T_i(\gamma, g, d)$, where i takes on values in the set $\{1, \dots, M\}$, and is asked to assess the presence of a discontinuity. Let the binary variable $R_i(T_i(\gamma, g, d))$ denote participant i 's discontinuity classification, which equals one if the participant reports a discontinuity at the policy threshold. Under random sampling, the following assumption holds:

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

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

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

Proposition 1 in Online Appendix A.1 states the distribution of $\hat{p}(\gamma, g, d)$ and shows the estimator to be unbiased and consistent as $M \rightarrow \infty$ 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 \mathcal{G} , which we formally define in Online Appendix A.2. While the infinite dimensionality of \mathcal{G} makes it difficult to characterize the distribution of DGPs, we think of the data used in empirical RD research as realizations when sampling from \mathcal{G} according to this distribution. 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 ($j = 1, \dots, J$) 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 \mathcal{G} .

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

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

the average of discontinuity classifications across the $M \cdot J$ classifications. Proposition 2 in Online Appendix A.1 states the distribution of $\hat{\bar{p}}(\gamma, d)$ and shows the estimator to be unbiased and consistent (as $J \rightarrow \infty$) 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 estimators would be closer in probability to the population parameters of interest). We henceforth refer to $\hat{p}(\gamma, g, 0)$ as the DGP-specific or g -specific type I error rate, and $\hat{\bar{p}}(\gamma, 0)$ as the average type I error rate (or simply the type I error rate). For a particular $d \neq 0$, we refer to $1 - \hat{p}(\gamma, g, d)$ as the DGP-specific or g -specific type II error rate, and $1 - \hat{\bar{p}}(\gamma, d)$ as the average type II error rate (or simply the type II error rate) at discontinuity d .

We can also define the type I and type II error probabilities of an econometric inference procedure based on a discontinuity estimator, $\hat{\theta}$, but with two adjustments. First, γ is no longer an argument in these expressions because we directly implement $\hat{\theta}$ on microdata W . Second, we need to specify the level of the testing procedure, 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 $\hat{p}(\gamma, d)$ as functions of d , which concisely summarize the type I and type II error probabilities of a graphical method. For brevity, we also use the term “power functions,” as opposed to “estimated power functions,” to refer to their empirical estimates. We study visual inference by comparing its power functions $\hat{p}(\gamma, d)$ across γ and against the corresponding power functions of various econometric inference procedures. We discuss the calculation of the standard errors on the differences between the visual and econometric power functions when we present our empirical results in Section 4, as its details depend on the design of our experiments.

The estimated power functions help shed light on best graphical practices. While the “optimal” graphical method depends in part on the optimality criterion we specify, the power functions provide key inputs for some common criteria that are well suited to this context. A simple criterion, as suggested by a referee, is how close the type I error rate is to 5 percent, the conventional threshold in econometric analysis. A second criterion quantifies the costs of type I and type II errors and uses the power function to consider explicitly the tradeoff between the two types of errors. We also consider a third criterion, which is adapted from Andrews and Shapiro (2021); instead of using the power function, this criterion relies on readers’ confidence in their discontinuity classification, which we can proxy using participants’ bonus scheme choice as discussed in Section 3.3 and Online Appendix D.1.2. A full discussion of the second and third criteria require formal decision-theoretic frameworks, and we leave it to Online Appendix A.3.

In summary, we have defined the type I and type II error rates for visual inference. The type I error rate is the fraction of continuous graphs which 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 probabilities of type I and type II errors which a randomly chosen person commits when classifying a graph generated from a representative DGP. These probabilities also help inform best graphical practices.

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. We discuss each of these treatments in detail below and provide graphical illustrations in Figure I.

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 hypothesis. Calonico, Cattaneo, and Titiunik (2015) propose two bin width selection algorithms based on different econometric criteria. The first, which is 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, Cattaneo, and Titiunik (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 $\frac{N}{\log(N)^2}$. The IMSE-optimal bin width therefore selects fewer bins, and hence has larger bin widths, than the MV algorithm (we describe these algorithms further in Online Appendix A.4). In addition to these two algorithms, Calonico, Cattaneo, and Titiunik (2015) provide an interpretation for any given number of bins as the output of a weighted IMSE-optimal algorithm. Applying the Lee and Lemieux (2010) algorithms to our datasets leads to bin numbers that tend to be between the IMSE and MV selectors and closer to those of the IMSE. Thus, we restrict our analysis to the visual inference properties of the IMSE and MV bin selectors.

Although the prevailing approach is to adopt evenly spaced bins, this method has drawbacks in that the resulting bins may contain vastly different numbers of observations, or even none at all.² This can happen when the distribution of the running variable is far from uniform. As a remedy, Calonico, Cattaneo, and Titiunik (2015) also propose quantile-spaced bins where each bin contains (approximately) the same number of observations. Both spacings support IMSE and MV bin selectors, and we test each of these

²In a literature review we conduct for current practices, 98% of the more than 100 studies we compile use evenly spaced bins. Our review includes RDD studies as well as studies that apply the regression kink design—RK design or RKD.

combinations.

Following suggestions by Imbens and Lemieux (2008) and Lee and Lemieux (2010), RD 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, Cattaneo, and Titiunik (2015) uses piecewise global quartic regressions on each side of the policy threshold. Of the 11 RDD papers on which we calibrate our DGPs, 10 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 conflicting recommendations (for example, for our first DGP, the Akaike information criterion selects a fifth-order polynomial, while the Bayesian information criteria and F -test they describe choose a zeroth-order polynomial).

The presence of a vertical line is designed to visually separate observations above and below the cutoff. We test all combinations of the fit line and vertical line treatments except for including the fit lines and excluding the 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. 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 and 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, 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: some of the replication data used for our DGP calibration are already subsets of a larger sample.

There are other 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.

3.2 Creation of Simulated Datasets and Graphs

We specify data-generating processes based on the actual data used in published research. We randomly sample 11 from a total of 110 empirical RD papers published in 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* between 1999 and 2017 that have replication data available to create our DGPs. We refer to these DGPs as 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 .³ We use the empirical distribution of the running variable from each of the 11 papers but normalize it to lie in $[-1, 1]$ by dividing X of each data point by the maximum $|X|$ on its side of the cutoff, with zero representing the location of that policy cutoff. We also remove the most extreme observations where $|X| > 0.99$, following Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014). For two papers which feature semidiscrete running variables, we add small amounts of normal noise to the running variable to match the regularity conditions from Calonico, Cattaneo, and Titiunik (2015). To create a continuous CEF, we fit global piecewise quintics (still following Imbens and Kalyanaraman, 2012 and Calonico, Cattaneo, and Titiunik, 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 (RMSE) of the piecewise quintic regression. We use the same number

³Out of the 11 studies, two plot residualized outcomes instead of the original Y to adjust for covariates, conceptually consistent with the covariate-adjusted RD estimation by Calonico et al. (2019) and the ideas of Angrist and Rokkanen (2015).

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 II, and Figure III illustrates the construction process. We describe the DGP creation process in full detail in Online Appendix B.

Because the outcomes from the 11 papers are measured in different units, we need to standardize the discontinuity levels and choose to specify discontinuity levels d as multiples of σ . Alternatively, we could specify d as multiples of the overall standard deviation of the outcome variable net of the discontinuity, a measure that also captures the variation due to the conditional expectation function besides the error term. Using this alternative measure turns out not to make a difference: the variance of the error term, σ^2 , dominates the variance of $E[Y|X]$ in all of our DGPs, with the ratio of the two ranging from 8 to 690.

As a multiple 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 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 this scale rather than a linear scale to generate more graphs with smaller discontinuities, which are harder to detect, 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 discontinuity 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.

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 from the respective studies. To do this, we adapt the lineup protocol from Buja et al. (2009) and Majumder, Hofmann, and Cook (2013), which uses visual inference to conduct hypothesis testing. 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. Under 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. We present two examples in Figure IV.

Based on visual testing among the authors, we cannot identify the graph from the true data for eight out of our 11 DGPs, which supports the idea that our DGPs approximate the original datasets well. For three DGPs, however, there is an obvious difference, as exemplified by DGP3 in the right panel of Figure IV. 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 the right panel of Figure IV, 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 inference procedures.

Another caveat of our DGP specification is that using global quintic regressions can lead to overfitting, the same issue that has brought forth the warning by Gelman and Imbens (2019) against using high-order global polynomial regressions to estimate RD treatment effects (see Pei et al., 2022 for related discussions on the order of local polynomial regressions). We acknowledge this potential drawback of using quintics as some of our graphs indeed feature high variation in the tails. That said, the lineup protocol we adapt offers a novel and transparent method to evaluate our DGP specifications, and we find our inability to distinguish the real data from those drawn from one of our DGPs in a majority of cases reassuring. To further assuage the concerns regarding our DGP specification, we carry out a supplemental phase of experiments to gauge the sensitivity of visual inference to alternative DGP specifications. In Online Appendix C, we demonstrate the remarkable robustness of our results to using local linear estimates as an alternative to model the CEF (and allowing for heteroskedasticity), which is much less likely to overfit.

3.3 Nonexpert Experiments

In our randomized experiments, we present nonexpert participants with binned scatter plots made from our DGPs and ask them to classify the graphs as having a discontinuity or not. We conduct five phases of computer-based experiments online through the Cornell University Johnson College’s Business Simulation Lab. Our subject pool consists of current and former Cornell students, Cornell staff, and nonstudent 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.⁴ 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 are attentive 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 explicitly tell participants that all, some, or none of the 11 graphs they classify may feature a discontinuity.

In each phase of the experiment, we present participants with a series of RD graphs using data generated as described in Section 3.2. 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. For each graph, we ask participants whether they believe there is a discontinuity at $x = 0$.

Because running an experiment with $2^5 = 32$ treatment arms is infeasible with our resources, we conduct our experiment in phases, testing only a few treatments in each phase. Table I details the timeline of the experiments and lists the graphical parameters we test and hold fixed for various experimental phases. In phase 1, we test both bin width and axis scaling options. In phase 2, we test bin widths and bin spacings. In phase 3, we test imposing fit lines and a vertical line at the treatment threshold. Based on the results from these three phases, in which only bin widths and fit lines have major impacts, phase 4 tests all four combinations of those two treatments together.

Participants receive a base pay of \$3 for being in the experiment. To stimulate participant engagement and elicit participants' confidence in their response, participants can choose for each graph they classify 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 classification. In Online Appendix D.1.2, we explore the implication of a participant's bonus choice and discuss how we use it—in addition to the type I

⁴The video tutorial is available at https://storage.googleapis.com/rd-video-tutorial/rd_video_tutorial.mp4.

and type II error rates—to evaluate graphical methods as mentioned in Section 2.

We do not give participants real-time feedback on the accuracy of their responses. Instead, we report total earnings and the final tally of correct classifications at the very end of the experiment after a short exit survey soliciting demographic information and comments.

The experiments are programmed in oTree (Chen, Schonger, and Wickens, 2016) and preregistered at the AEA RCT registry (Korting et al., 2019a) and the Center for Open Science’s OSF platform (Korting et al., 2019b). The study takes participants approximately 15 minutes to complete.

3.4 Expert Study

In addition to our nonexpert experiment, we conduct a study with researchers in economics and related fields who work on topics that often employ RDDs. We collect data at three technical social science seminars and online by contacting randomly selected members of the NBER in applied microeconomic fields (aging, children, development, education, health, health care, industrial organization, labor, and public) and IZA fellows and affiliates. After removing 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.

This expert study allows us to answer two questions. First, how do classification accuracy and the impacts of graphical techniques differ between experts and nonexperts? And second, can experts correctly predict which graphing options perform best for our nonexpert sample? This second question speaks to experts’ ability to predict which visualization choices are best 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 nonexpert performance, DellaVigna and Pope (2018) find that economic experts are better than nonexperts at estimating the effect of alternative incentive schemes on performance in a real-effort task, but perform similarly to nonexperts in terms of a simple ranking of incentive schemes.

Our expert study consists of two parts. The first is similar in structure to the nonexpert experiment. Participants see a series of RD graphs and are asked to classify them by whether they have a discontinuity. To assess the accuracy of point estimates in addition to binary classifications of discontinuities, we also ask participants 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 equally spaced bins, no fit lines, default axis scaling, and a vertical line at the treatment threshold. All expert graphs use small bins, except for one seminar where participants see large bins. Four randomly selected participants receive a base payment of \$450 plus a bonus payment of \$50 per correct discontinuity classification. The bonus payment 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 magnitudes: 0, 0.54σ , and 1.5σ . At each magnitude, we present four graphs, one for each combination of bin width and fit lines, in a random order using the same underlying data from the DGP where visual inference performs most closely to the average across all 11 DGPs. At each magnitude, we ask the experts to indicate which of the four treatment options they prefer and which they believe perform best and worst in our nonexpert sample. We evaluate the experts' predictions about nonexpert performances using phase 4 of the nonexpert experiment, which tests these four treatment permutations simultaneously.

4 Results

4.1 Nonexpert Experiment Results and Graphical Method Recommendation

For each combination of graphical parameters in all phases of the experiment, we compute power functions based on participants' classifications of whether graphs feature 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 $\hat{p}(\gamma, d)$.

The intercept of a power function indicates the type I error rate 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 one minus the type II error rate when the DGP is chosen uniformly randomly from the 11 possibilities. A desirable inference method has a small intercept before quickly rising to achieve high power. Plots of the overall power functions for each phase are shown in Figure V, and Figure A.1 plots the corresponding DGP-specific power functions. 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 effects on visual inference are in Tables A.1-A.5. 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 1 of the experiment 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, large bins have a significantly higher type I error rates relative to small bins. While both small bin treatments lead to a type I error rate of approximately 5%, the large bins have a type I error rate of around 20% to 25%. The large bins have a type II error advantage over the small bins, which is related to their higher type I error rate, but the power functions converge 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 2, 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 1 and 2. With this design, we can gauge the stability of visual inference across different samples from the nonexpert 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 evenly spaced and quantile-spaced bins perform very similarly. We conclude that bin spacing has a small or null effect on visual inference for the DGPs we test.

Phase 3 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 type I error rates 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 nonexpert performance on the interaction of these two treatments and run a final phase of experiments, phase 4, 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 bins and fit lines has a type I error rate of only 0.052 in phase 4 but 0.175 in phase 3. This finding suggests that we cannot conclude that fit lines unequivocally result in an increase in type I errors, but that they do add uncertainty to visual inference.

We provide additional results in Online Appendices F and G. Online Appendix F 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. Online Appendix G examines how large the t -statistics need to be for readers to detect a discontinuity visually.

Based on our experimental results, we recommend the graphical method that uses small bins, no fit lines, even spacing, default y-axis scaling, and a vertical line at the policy threshold as a sensible default for generating RD graphs. It has an associated type I error rate close to 5 percent in all four phases of the experiment. As we document in Online Appendix E, it also performs well under the two additional criteria mentioned in Section 2. Of the five graphical parameters, bin spacing, y-axis scaling, and the presence of the vertical line do not appear to matter much, allowing researchers to use reasonable discretion. The other two parameters are much more important, and the use of small bins and no fit lines appears to be key for good visual inference performance.

Finally, we emphasize that our recommendation is not intended as a doctrine that practitioners must abide by. We are limited to the 11 DGPs we test. In addition, as described in Online Appendix A.4, there is an ad hoc element to the construction of small bins. In fact, we use quantile-spaced large bins in the regression kink design experiments in a previous working paper Korting et al. (2020), where the sample sizes are much larger than for our RD DGPs. In this regard, we view our recommended graphical method as a reasonable starting point based on the best evidence we have. An equally important takeaway is the value in documenting the robustness of graphical evidence given our finding of divergent visual inferences under commonly used methods.

4.2 Expert Study Results

We show most expert participants (95 out of 143) graphs generated with our preferred method as discussed above. Figure VI plots the expert power functions against those of the nonexperts who saw the same graphs. When comparing expert and nonexpert performances, we use solid (hollow) markers to indicate that the point is (not) statistically significantly different from the reference curve, and present plots of the corresponding 95% confidence intervals (created here with the large sample approximation described at the end of Online Appendix A.1 and by assuming independence between the experts and nonexperts) in Figure A.30. The two groups perform similarly, with experts having a slightly higher type I error rate (approximately 8% to the nonexpert 5%) and a slightly lower type II error rate. The only statistically significant differences are

for the experts' marginally lower type II error rates 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. The two groups again perform similarly, and the expert and nonexpert power functions are not statistically significantly different anywhere. Both groups have type I error rates well above their corresponding small bin rates. Like nonexperts, experts do worse when viewing graphs constructed with large bins.

4.2.1 Expert Preferences and Predicting Nonexpert Performance

We present experts' preferences and their beliefs about nonexpert performance across the four considered treatments in Figure VII. 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 nonexperts tend to mirror their preferences. By a large margin, experts believe small bins with fit lines to be the most efficacious treatment for nonexperts at all discontinuity magnitudes. Conversely, most experts view large bins without fit lines least favorably in the context of nonexpert performance.

Comparing the expert predictions to our experimental data from phase 4, we find substantial discordance for the effects of bin width choice on nonexpert classification accuracy. The best- and worst-performing treatments at each discontinuity magnitude have + and - signs, respectively, in Figure VII. The actual power functions are shown in Figure V (Online Appendix Figure A.3 shows the power functions based only on DGP9, the DGP used in the example graphs shown to experts in the second part of the expert study). While a majority of experts correctly identifies the bin width treatment with lowest type I error rates (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 lowest type II error rates 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 nonexpert performance on bin width than on fit lines. We also find expert predictions to be similar whether their own visual inference performance is above or below the median.

4.3 Visual versus Econometric Inference

Here we compare the performances of visual inference from our small-bin expert sample with various econometric RD procedures. We present both the overall power functions and the difference between visual and econometric inferences for each econometric procedure. 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.

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 the corresponding 5% test compared with human performance is presented in the left panel of Figure VIII. We again use solid (hollow) markers to indicate that the difference to the comparison power function is (not) statistically significant, and provide plots of the differences in Online Appendix Figure A.31. We additionally include in Table II type I and II error rates (the latter at both $|d| = 0.324\sigma$ and averaged across nonzero discontinuities) for visual and econometric inferences.

Next, we implement the IK, CCT, and AK inference procedures, again plotting the corresponding power functions in the left panel of Figure VIII. All three procedures build upon local linear regressions but take different approaches to conducting inference. We introduce them here briefly before discussing the performance of each. We provide a more detailed review of the methods in Online Appendix H.

Strictly speaking, Imbens and Kalyanaraman (2012) do not study inference but propose an MSE-optimal bandwidth selector, the IK bandwidth. The IK inference procedure we refer to is the "conventional" (terminology from Calonico, Cattaneo, and Titiunik, 2014) inference procedure practitioners typically implement in conjunction with the IK bandwidth in which the asymptotic bias is ignored. As seen in the left panel of Figure VIII, the IK procedure achieves even lower type II error rates than the piecewise quintic estimator, and is significantly better than visual inference at detecting discontinuities up to 1.5σ . But with this advantage in type II error rate comes a significant disadvantage in type I error rate. When there is truly no discontinuity, the estimator still rejects the null hypothesis in 22.6% of datasets.

Unlike IK, the CCT procedure by Calonico, Cattaneo, and Titiunik (2014) directly estimates the asymptotic bias, centers the confidence interval at the bias-corrected estimate, and adjusts the width of the confidence interval to account for the uncertainty in the bias estimate. CCT also generalize IK and propose a new

class of MSE-optimal bandwidth selectors, which are implemented in the Stata package `rdrobust` along with the inference procedure.⁵ Note that although CCT's type I error rate is approximately 12.5%, as seen in the left panel of Figure VIII, this could be a small sample problem. As mentioned in Section 3.4, we effectively have 88 datasets modulo the discontinuity level for the expert study. In a separate Monte Carlo simulation with 1,000 draws, the type I error rate 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 datasets the experts saw. The CCT inference procedure achieves lower type II error rates, enjoying a significant 15 to 20 percentage point advantage over expert visual inference at intermediate discontinuity levels.⁶

Finally, the AK inference procedure by Armstrong and Kolesár (2018) adapts Donoho (1994) and produces asymptotically valid and minimax (near-)optimal confidence intervals over a class of conditional expectation functions with a bound—loosely speaking—on the second-derivative magnitudes just above and just below the cutoff. This bound informs the worst-case bias of a local linear estimator over the class of functions, and AK corrects for this bias in their procedure (unlike CCT, AK's bias correction is nonrandom). Using the default rule-of-thumb method in the `RDHonest` R package to estimate the tuning parameter, the AK inference has a type I error rate of approximately 6%, and the power function is very close to that of the experts.⁷ Given the minimax optimality of AK, the comparable performance of visual inference is remarkable. However, it is worth emphasizing that although they have approximately the same average type I error rate, AK offers a theoretical guarantee to control the (asymptotic) type I error rate for all DGPs in the Taylor class while visual inference does not.

The IK and CCT inference procedures at the 5% level exhibit higher type I and lower type II error rates

⁵While 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 Online Appendix Figure A.4, using these bandwidths reduces the excess type I error rate relative to visual inference.

⁶One may be concerned that visual inference's lower type I error rate is a consequence of our experimental design where the majority (9 out of 11) of graphs feature a discontinuity. If subjects speculate that only about half the graphs feature a discontinuity, our type-I-error-rate result is biased in favor of visual inference. As mentioned in Section 3.3, 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 against this bias in Online Appendix F.3, where we test dynamic visual inference.

⁷Armstrong 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. Though not presented here, we find the corresponding power functions to be similar. 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 the corresponding inference procedure performs similarly to Armstrong and Kolesár (2018) in simulations by Pei et al. (2022) in their 2018 working paper version and can be computationally demanding, we do not implement the Imbens and Wager (2019) procedure here.

than visual inference. We can circumvent this tradeoff by adjusting their type I error rate to the level of visual inference: we search for alternative critical t -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 type-I-error-rate-adjusted inference procedures in the right panel of Figure VIII, with the differences between these procedures and expert visual inference in Online Appendix Figure A.32. Despite the extent of their differences in type I error rates relative to visual inference, both econometric procedures' type II error rates only increase by around 5–10 percentage points from this adjustment and are still significantly lower than those of visual inference at moderate discontinuities. However, we again caution that this result is specific to the sample of the 11 DGPs we consider; the type-I-error-rate adjusted IK and CCT procedures are not guaranteed to have correct coverage over the broader Taylor class as AK does.

We provide two sets of additional results in Online Appendix H. First, to investigate the mechanisms that underlie the econometric methods' performances, we impose our knowledge of the DGP when implementing them. We find that the driving force of IK's high type I error rate appears to be the noisy estimates that feed into the optimal bandwidth formula. When we use the theoretical optimal bandwidth, the corresponding type I error rate is lower and statistically indistinguishable from visual inference's.

Second, we venture beyond binary classifications of a discontinuity and compare the accuracy of discontinuity point estimates across visual and econometric approaches. We find that a simple method like IK attains a lower RMSE than visual inference for each of the 11 DGPs. We conclude that visual inference is less competitive at estimating discontinuity magnitudes than at identifying their existence.

4.3.1 The Complementarity of Visual and Econometric Inferences

Thus far, we have studied the “marginal” power functions for visual and econometric inferences. In this section, we use the joint distribution of visual and econometric discontinuity tests to explore their complementarity and evaluate the performance of a simple combined visual-econometric inference procedure.

First, we examine the joint distribution of visual and econometric inferences to see whether they tend to agree on the same data. For each discontinuity magnitude, we characterize the joint distribution of visual and econometric classifications in the form of a two-by-two contingency table. We conduct Fisher's exact test for independence and present one-sided p -values in Table III by discontinuity magnitude and for each econometric method. We report one-sided p -values because two-sided p -values are method-dependent due

to the ambiguity in classifying contingency tables as extreme in the opposite direction (see, e.g., Agresti, 1992).⁸

Two patterns emerge from Table III. The first row shows no strong support for an association between visual and econometric classifications when the true discontinuity is zero. But when the true discontinuity is nonzero (rows two through six), there appears to be strong evidence in support of association (though not reported in Table III, 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, which highlights the complementarity of visual and econometric inferences.

Second, we illustrate this complementarity more concretely by studying the performance of a particular combined visual-econometric inference procedure. It infers a discontinuity if and only if both the visual and econometric procedures reject the null hypothesis of no discontinuity. As a referee points out, many researchers may already use this procedure informally when reading or writing RD papers.

We plot the resulting power functions in the top panel of Figure IX along with the power function of the AK procedure for comparison. Because of the lack of dependence between visual and econometric classifications when $d = 0$, the combined inferences achieve lower type I error rates than either of the individual inference types. On the other hand, the same mechanism pushes type II error rates higher, but the positive associations between the classifications when $d \neq 0$ help to limit their increase.

In fact, the power function of the combined visual-IK inference procedure is fairly close to that of AK. The bottom panel of Figure IX presents the difference between the two. Despite the limitations of the IK procedure, with a type I error rate above 20% as shown before, the IK-expert hybrid has a type I error rate of 2.6% while not performing statistically significantly differently from AK at any of the nonzero discontinuity levels. This finding helps to explain the enduring credibility of RDDs despite potential issues with the econometric inference method used prior to CCT and AK. It also suggests that the *de facto* type I error probability may be lower than the nominal level if researchers informally combine different statistical evidence instead of relying on a single econometric inference result, a point that deserves attention in future research.

⁸In principle, we could also report correlations between inferences, but they may be hard to interpret. Because our classification variables are binary, the maximum of the correlation measure depends on the marginal distributions of the classifications. For example, if the probabilities of rejecting no discontinuity are different between two inference procedures, then their correlation is strictly less than one. Because the (marginal) classification probabilities vary across discontinuity levels and by methods, it is difficult to compare the correlation measures.

5 Conclusion

This paper studies visual inference and graphical representation in RD designs via crowdsourcing. Through a series of experiments and studies that recruit both nonexpert and expert participants, we provide answers to two sets of questions. First, how do graphical representation techniques affect visual inference and which technique should practitioners use? And second, when presented with well constructed graphs, how does visual inference perform compared to common econometric inference procedures in RDDs?

To answer the first set of questions, we experimentally assess how five graphical parameters impact visual inference accuracy. We find that generating graphs with the Calonico, Cattaneo, and Titiunik (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. Imposing fit lines can have a similar effect as using large bins, confirming the worries by Cattaneo and Titiunik (2021a,b). We recommend the graphical method of using small bins and no fit lines as a sensible starting point in practice, which we show performs well under several criteria. Bin spacing, a vertical line at the policy threshold, and y-axis scaling have little effect, implying that researchers can adhere to reasonable preferences.

For the second set of questions, we find that visual inference performs competitively on graphs constructed with the recommended method. It achieves a lower type I error rate than econometric inference at the 5% level based on the Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) methods (the difference between the visual and CCT type I error rates is not statistically significant), though the two econometric inference procedures offer considerable type-II-error advantages. The performance of visual inference is very similar to that based on the procedure suggested by Armstrong and Kolesár (2018). Furthermore, visual and econometric inferences appear to be complementary. Through the analysis of the joint distribution of visual and econometric tests we find that, while they commit similar type II errors, there does not appear to be a strong association in their type I errors.

Our study is subject to several important limitations. The first is the restricted set of parameters we are able to test experimentally. As mentioned earlier, we do not impose fit lines with confidence intervals in our graphs, which researchers sometimes do, due to the difficulty in explaining it to nonexpert 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.

Second, our results are based on a specific set of DGPs. For example, while the bin spacing choice—equally spaced or quantile spaced—appears immaterial in our 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 often far lower than our 11, and those DGPs sometimes bear no resemblance to real-world data. In addition, we test the validity of our simulated datasets by adapting the lineup protocol from Majumder, Hofmann, and Cook (2013) to assess the degree to which our DGPs approximate the original data, and we document the robustness of our experimental results to alternative DGP specifications.

And third, the mechanism of RD visual inference remains elusive. In a previous working paper (Korting et al., 2020), we reported the results from an eyetracking study, in which we sought to identify eyegaze patterns (e.g., “visual bandwidths”) that robustly predict visual inference success. Had predictive patterns emerged from the eyetracking study, we would have followed up with additional experiments, in which we instruct a random subset of the participants to focus their visual attention according to our finding. But we were not able to identify predictive ocular patterns and could only conclude that the processing of visual signals, as opposed to where in a graph participants looked, drove visual inference success. A next step toward better understanding the mechanism is to systematically study the types of DGPs for which visual inference performs well and poorly.

These limitations notwithstanding, our study answers the call by Leek and Peng (2015) to provide empirical evidence on best practices in data analysis, and our approach can find applications in other important areas. We have conducted analogous experiments to study visual inference and graphical representations in RKDs using DGPs based on Card, Lee, and Pei (2009), Card et al. (2015a), and Card et al. (2015b) (Ganong and Jäger, 2018 also discuss RK visual inference, albeit informally). Interested readers can consult our previous working paper (Korting et al., 2020) for our nuanced findings. Another related topic to study follows the recent work by Cattaneo et al. (2019), 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. A third related topic is structural breaks in time series econometrics, in which graphs serve practically the same purpose as those

in RDDs. Finally, within time series econometrics, studying visual inference for unit root/stationarity analysis may also be promising.⁹ 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?

Author Affiliations

University of Delaware

U.S. Census Bureau

Columbia University

Cornell University and IZA

University of Waterloo

Supplementary Material

An Online Appendix for this article can be found at The Quarterly Journal of Economics online.

Data Availability

Data and code replicating the tables and figures in this article can be found in Korting, Lieberman, Matsudaira, Pei, and Shen (2023) in the Harvard Dataverse, <https://doi.org/10.7910/DVN/RAHKNJ>.

⁹In 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, Alan. 1992. "A Survey of Exact Inference for Contingency Tables." *Statistical Science* 7 (1):131–153.
- Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro. 2020. "Transparency in Structural Research." *Journal of Business & Economic Statistics* 38 (4):711–722.
- Andrews, Isaiah and Jesse M Shapiro. 2021. "A Model of Scientific Communication." *Econometrica* 89 (5):2117–2142.
- Angrist, Joshua D. and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114 (2):533–575.
- Angrist, Joshua D. and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff." *Journal of the American Statistical Association* 110 (512):1331–1344.
- Armstrong, Timothy B and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86:655–683.
- . 2020. "Simple and Honest Confidence Intervals in Nonparametric Regression." *Quantitative Economics* 11 (1):1–39.
- Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1906):4361–4383.
- Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell. 2020. "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs." *The Econometrics Journal* 23 (2):192–210.
- Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *Review of Economics and Statistics* 101 (3):442–451.
- Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.
- . 2015. "Optimal Data-Driven Regression Discontinuity Plots." *Journal of the American Statistical Association* 110 (512):1753–1769.
- Camerer, Colin F. and Eric 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* 342:195–217.
- Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference With Multiway Clustering." *Journal of Business & Economic Statistics* 29 (2):238–249.
- Card, David, Andrew Johnston, Pauline Leung, Alexandre Mas, and Zhuan 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 (5):126–130.

- Card, David, David S. Lee, and Zhuan Pei. 2009. "Quasi-Experimental Identification and Estimation in the Regression Kink Design." Princeton University Industrial Relations Section Working Paper 553.
- Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015b. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83 (6):2453–2483.
- Casella, George and Roger L Berger. 2002. *Statistical Inference*. Cengage Learning; 2nd edition.
- Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2019. "On Binscatter." *arXiv preprint arXiv:1902.09608*.
- Cattaneo, Matias D, Nicolás Idrobo, and Rocío Titiunik. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.
- Cattaneo, Matias D and Rocío Titiunik. 2021a. "Causal Inference Using Synthetic Controls and Regression Discontinuity Designs." <https://www.nber.org/lecture/summer-institute-2021-methods-lectures-causal-inference-using-synthetic-controls-and-regression>. National Bureau of Economic Research Summer Institute 2021 Methods Lectures, last accessed 2021-11-09.
- Cattaneo, Matias D and Rocio Titiunik. 2021b. "Regression Discontinuity Designs." *arXiv preprint arXiv:2108.09400*.
- Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.
- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* 126 (4):1593–1660.
- Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *The Quarterly Journal of Economics* 129 (4):1553–1623.
- Cleveland, William S., Persi Diaconis, and Robert McGill. 1982. "Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased." *Science* 216 (4550):1138–1141.
- Cleveland, William S. and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387):531–554.
- Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data are Changing Economics: Mining Text to Track Methods." *AEA Papers and Proceedings* 110:42–48.
- Della Vigna, Stefano and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6):2410–2456.
- Donoho, David L. 1994. "Statistical Estimation and Optimal Recovery." *The Annals of Statistics* :238–270.
- Dynarski, Susan. 2014. "What We Mean when We Say Student Debt is Bad." *New York Times* URL <https://www.nytimes.com/2014/08/09/upshot/what-we-mean-when-we-say-student-debt-is-bad.html>.
- Bells, Walter Crosby. 1926. "The Relative Merits of Circles and Bars for Representing Component Parts." *Journal of the American Statistical Association* 21 (154):119–132.

- Ganong, Peter and Simon Jäger. 2018. “A Permutation Test for the Regression Kink Design.” *Journal of the American Statistical Association* 113 (522):494–504.
- Gelman, Andrew and Guido Imbens. 2019. “Why high-order polynomials should not be used in regression discontinuity designs.” *Journal of Business & Economic Statistics* 37 (3):447–456.
- Hegarty, Mary, Matt S Canham, and Sara 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 (1):37.
- Imbens, Guido and Karthik Kalyanaraman. 2012. “Optimal Bandwidth Choice for the Regression Discontinuity Estimator.” *The Review of Economic Studies* 79 (3):933–959.
- Imbens, Guido and Stefan Wager. 2019. “Optimized Regression Discontinuity Designs.” *Review of Economics and Statistics* 101 (2):264–278.
- Imbens, Guido W. and Thomas Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics* 142:615–635.
- Korting, Christina, Carl Lieberman, Jordan Matsudaira, Zhuan Pei, and Yi 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.
- . 2020. “Visual Inference and Graphical Representation in Regression Discontinuity Designs.” Princeton University Industrial Relations Section Working Paper 638.
- . 2023. “Replication Data for: ‘Visual Inference and Graphical Representation in Regression Discontinuity Designs’.” Harvard Dataverse, <https://doi.org/10.7910/DVN/RAHKNJ>.
- Kriz, Sarah and Mary Hegarty. 2007. “Top-Down and Bottom-Up Influences on Learning from Animations.” *International Journal of Human-Computer Studies* 65 (11):911–930.
- Lee, David S. and Thomas Lemieux. 2010. “Regression Discontinuity Designs in Economics.” *Journal of Economic Literature* 48 (2):281–355.
- Leek, Jeffrey T. and Roger D. Peng. 2015. “Statistics: P Values are Just the Tip of the Iceberg.” *Nature News* 520 (7549):612.
- Li, Housen, Axel Munk, Hannes Sieling, and Guenther Walther. 2020. “The Essential Histogram.” *Biometrika* 107 (2):347–364.
- Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. “Validation of Visual Statistical Inference, Applied to Linear Models.” *Journal of the American Statistical Association* 108 (503):942–956.
- Pei, Zhuan, David S. Lee, David Card, and Andrea Weber. 2022. “Local Polynomial Order in Regression Discontinuity Designs.” *Journal of Business & Economic Statistics* 40 (3):1259–1267.
- Rosen, Rebecca J. 2015. “Slight Changes in Yelp Ratings can Mean Huge Losses for Small Businesses.” *The Atlantic* URL <https://www.theatlantic.com/technology/archive/2012/09/slight-changes-in-yelp-ratings-can-mean-huge-losses-for-small-businesses/261943/>.
- Rosenholtz, Ruth, Yuanzhen Li, and Lisa Nakano. 2007. “Measuring Visual Clutter.” *Journal of Vision* 7 (2):17–17.

- Sanders, Michael, Freddie Mitchell, and Aisling Ni 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.
- Schwartzstein, Joshua and Adi Sunderam. 2021. “Using Models to Persuade.” *American Economic Review* 111 (1):276–323.
- Shen, Yi and Tony S. Wirjanto. 2019. “Stationarity as a Path Property.” *Probability and Mathematical Statistics* 39:403–422.
- Sides, John. 2015. “How to Get Young People to Vote? Register Them Before they Turn 18.” *The Washington Post*.
- Stock, James H. and Mark W. Watson. 2011. *Introduction to Econometrics*, 3rd edition. Pearson.
- Watson, Geoffrey S. 1964. “Smooth Regression Analysis.” *Sankhyā: The Indian Journal of Statistics, Series A* 26 (4):359–372.
- Wilkinson, Leland. 2013. *The Grammar of Graphics*. Springer Science & Business Media.
- Wolfe, Jeremy M and Todd S Horowitz. 2004. “What Attributes Guide the Deployment of Visual Attention and How Do They Do it?” *Nature Reviews Neuroscience* 5 (6):495–501.

Tables

Table I
Timeline of Experiments and Graphical Parameters Tested

Phase	Holding Fixed	Treatments	Date	# Recruited [# Completions (Rate)]
Main Phases				
1	bin spacing: ES fit lines: no vertical line: yes	bin width: large vs small # axis scaling: normal vs large	Nov. 13–16, 2018	4 × 88 = 352 [330 (94%)]
2	axis scaling: default fit lines: no vertical line: yes	bin width: large vs small # bin spacing: ES vs QS	Feb. 11–12, 2019	4 × 88 = 352 [325 (92%)]
3	bin width: small bin spacing: ES axis scaling: default	fit lines: no; vertical line: yes fit lines: no; vertical line: no fit lines: yes; vertical line: yes	Feb. 27, 2019	3 × 88 = 264 [248 (94%)]
4	bin spacing: ES axis scaling: default vertical line: yes	bin width: large vs small # fit lines: yes vs no	Oct. 28–29, 2019	4 × 88 = 352 [340 (97%)]
Supplemental Phase				
5	bin width: small fit lines: no bin spacing: ES axis scaling: default vertical line: yes	global quintic vs local linear specification # homoskedastic vs heteroskedastic error	Mar. 10–11, 2021	4 × 88 = 352 [339 (96%)]

Notes: In our four main experimental phases, we test the effects of:

- the bin width selector (we choose two bin width algorithms from Calonico, Cattaneo, and Titiunik, 2015: the first, called *large* above, minimizes the integrated mean squared error of the bin-average estimators of the conditional expectation function and results in fewer, larger bins; the second, called *small* above, aims to approximate the variability of the underlying data and results in more, smaller bins);
- bin spacing (evenly spaced, called *ES* above, and quantile spaced, called *QS* above);
- parametric fit lines;
- a vertical line at the policy threshold; and
- y-axis scaling (the default output from Stata 14, called *normal* above, and an increased scale created by recording the range of the y-variable from the default graph and increasing the bounds by 50% of the original range in each direction, called *large* above).

In the supplemental phase, we test the sensitivity of visual inference to alternative specifications of the data-generating processes.

Table II
Type I (False Positive) and Type II (False Negative) Error Rates for Expert and Econometric Inferences

Inference Type	Type I Error Rate	Type II Error Rate: $ d = 0.324\sigma$	Average Type II Error Rate: $d \neq 0$
Experts	0.079	0.537	0.336
Piecewise Quintic	0.068 [.785]	0.395 [.003]	0.260 [.000]
IK	0.226 [.005]	0.216 [.000]	0.145 [.000]
CCT	0.132 [.252]	0.342 [.000]	0.219 [.000]
AK	0.058 [.585]	0.500 [.429]	0.321 [.400]

Notes: The second column shows the type I (false positive) error rate. The third column shows the type II (false negative) error rate for a discontinuity of $|d| = 0.324\sigma$. The fourth column shows the average type II error rate when $d \neq 0$, weighting all discontinuity magnitudes equally. We present p -values for the difference between experts and each estimator in brackets. They are based on two-way cluster-robust standard errors computed via a stacked regression where we account for the potential correlation between visual and econometric inferences at the data set level (there are 88 data sets in total) and in visual inferences for the same individual across graphs—see Online Appendix H.1 for details. IK inference is based on a local linear estimator using the IK bandwidth (Imbens and Kalyanaraman, 2012). CCT is the default RDD inference procedure from CCT’s `rdrobust` (Calonico, Cattaneo, and Titiunik, 2014). AK uses the `RDHonest` procedure with the rule-of-thumb bound on each DGP’s second derivative (Armstrong and Kolesár, 2018).

Table III
Fisher's Exact Test of Association: Expert Visual vs Econometric Inferences p -values (One-Sided)

Discontinuity $ d $	Estimator for Econometric Inference			
	PQ	IK	CCT	AK
0	.727	.231	.616	.394
0.1944σ	.000	.000	.000	.000
0.324σ	.000	.000	.000	.000
0.54σ	.000	.000	.001	.000
0.9σ	.073	.	.042	.134
1.5σ

Notes: Missing values indicate that a test always rejects the null hypothesis, in which case the p -value cannot be computed. PQ uses a correctly specified regression model with global piecewise quintics above and below the treatment threshold and assuming homoskedasticity. IK inference is based on a local linear estimator using the IK bandwidth (Imbens and Kalyanaraman, 2012). CCT is the default RDD inference procedure from CCT's `rdrobust` (Calonico, Cattaneo, and Titiunik, 2014). AK uses the `RDHonest` procedure with the rule-of-thumb bound on each DGP's second derivative (Armstrong and Kolesár, 2018).

Figures

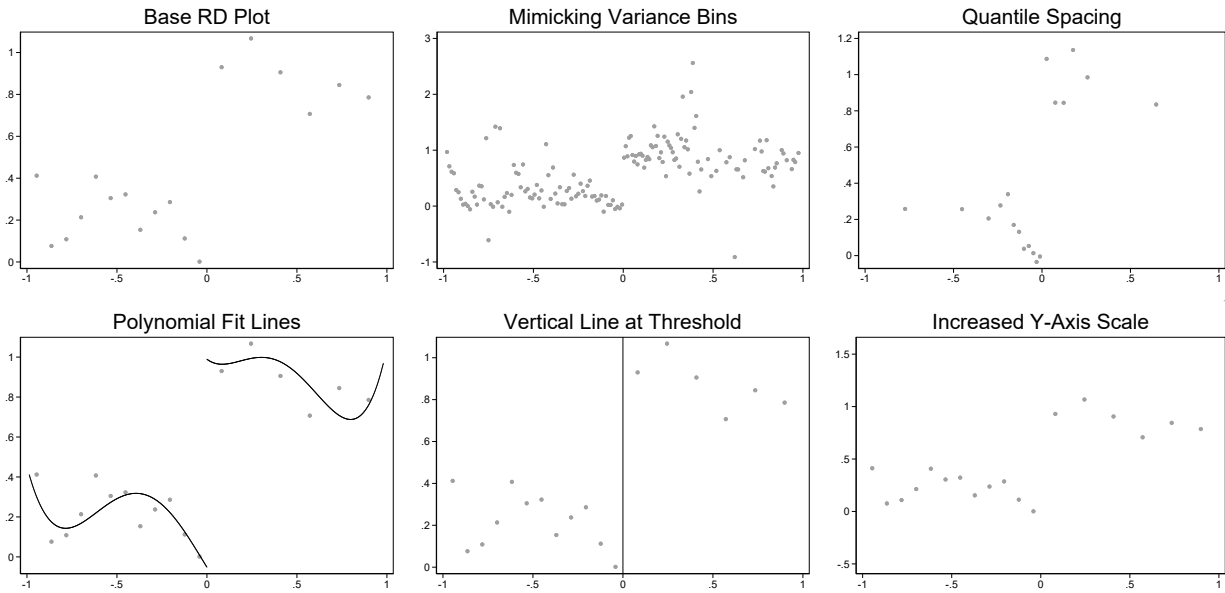


Figure I
Illustration of Graphical Parameters Tested

Notes: Plots are based on the original data from DGP9. The “Base RD Plot” uses evenly spaced IMSE-optimal (or “large”) bins and Stata 14’s default axis scaling. The graph labeled “Mimicking Variance Bins” uses mimicking-variance (or “small”) bins and maintains equal bin spacing and default axis scaling. The “Quantile Spacing” graph uses quantile-spaced large bins and default axis scaling. The three graphs in the second row present the same binned points as the base plot but with different graphical options: the first imposes parametric polynomial fit lines, the second adds a vertical line indicating the policy threshold at zero, and the third uses a y-axis scaling that is twice the Stata 14 default. The bin selectors and spacings—large, small, equally spaced, and quantile spaced—come from the study by Calonico, Cattaneo, and Titiunik (2015).

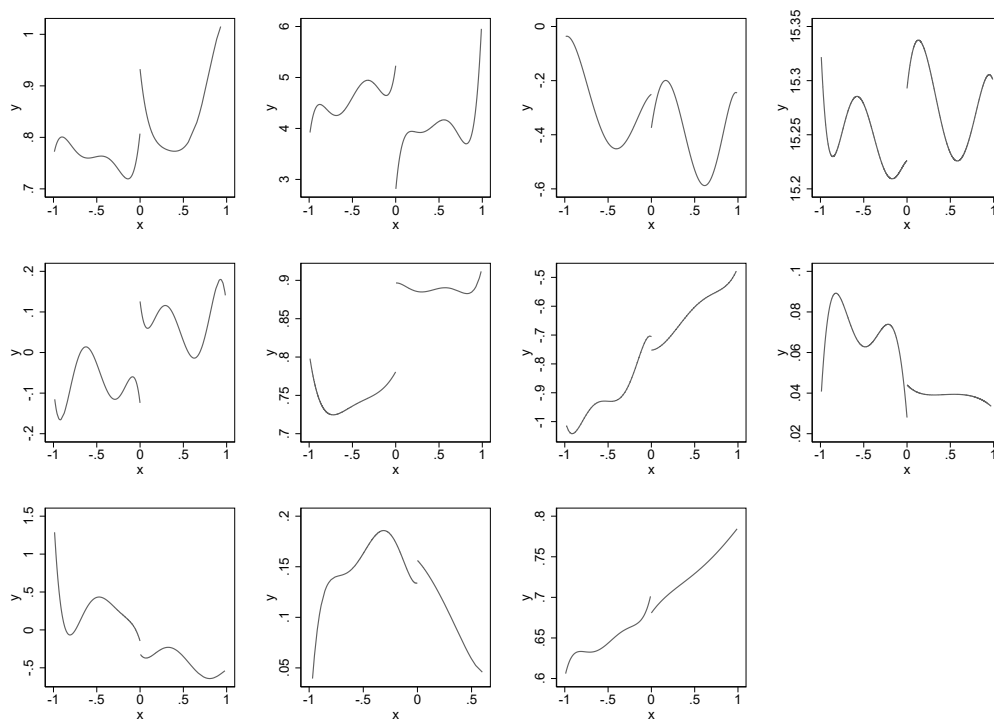


Figure II
Conditional Expectation Functions DGP1–DGP11

Notes: Each panel represents the plot of the conditional expectation function for DGP1 through DGP11. These functions are obtained by fitting piecewise global quintics to the original microdata after normalizing the support of the running variable and trimming the tails, as described in Section 3.2.

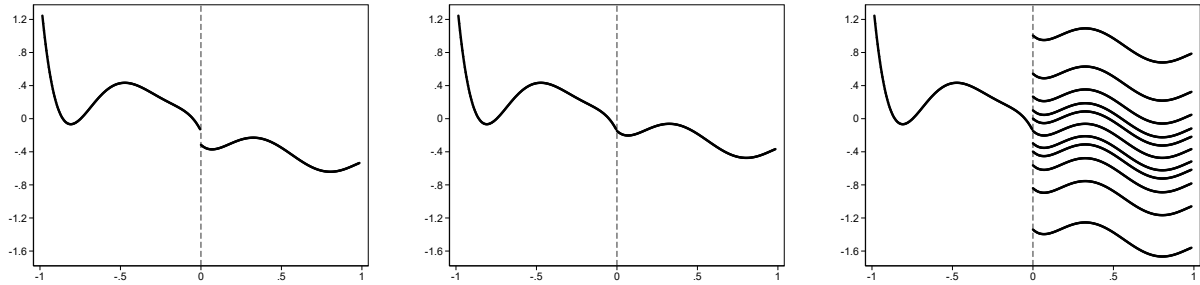


Figure III
Creation of Conditional Expectation Functions, DGP9

Notes: The leftmost figure plots the piecewise quintic CEF fitted to the original microdata underlying DGP9 after normalizing the support of the running variable and trimming the tails, as described in Section 3.2. 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.

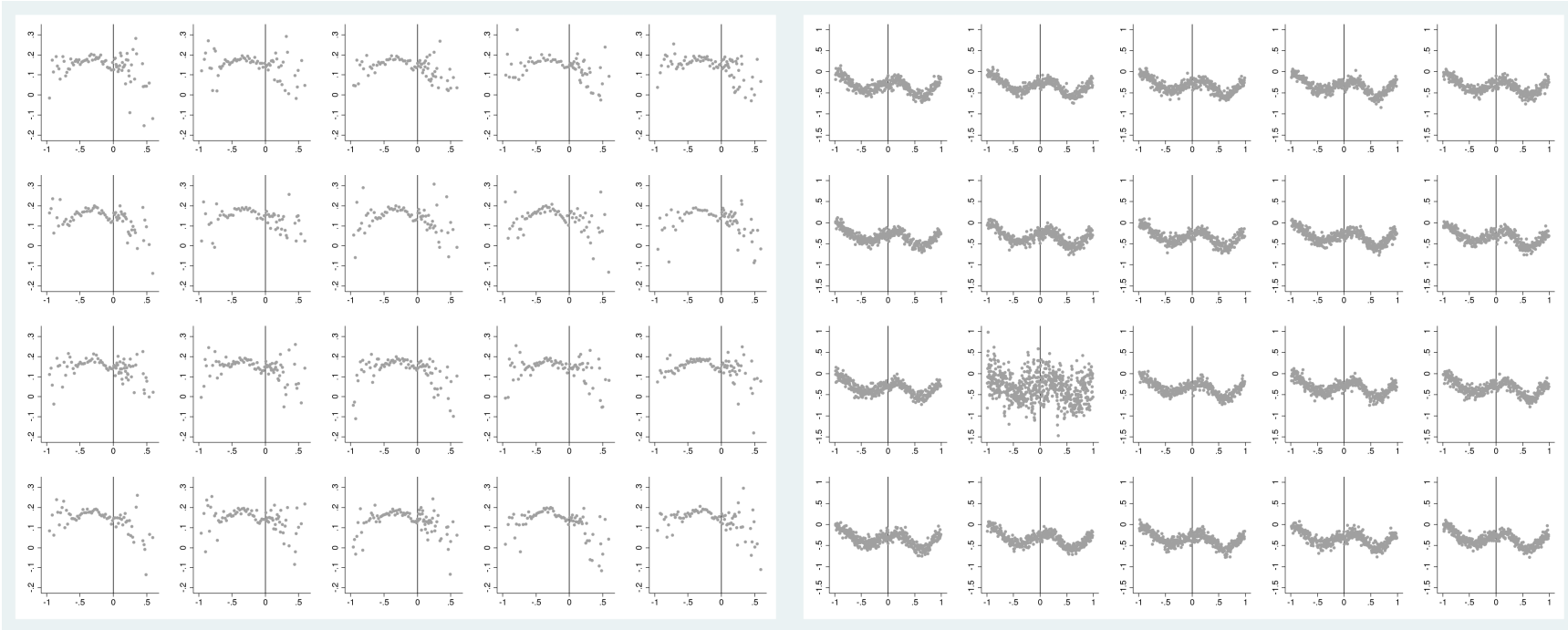


Figure IV
Lineup Protocol Graph Examples: DGP10 and DGP3

Notes: One of the 20 graphs for each lineup protocol is produced from the real data. The other 19 are produced from simulated data drawn from the DGP calibrated to the real data. For DGP10 on the left, the graph produced from data used in the original paper is in row $-3 \cdot 2 + 7$ and column $\sqrt{4+5} - 2$ (using conventional matrix index notation), while the remaining graphs are generated from our specified DGP (we follow Majumder, Hofmann, and Cook, 2013 and use simple arithmetic to indicate the graph location, so that readers do not accidentally see the answer before reaching their own conclusion). For DGP3 on the right, the graph made with the original data is in row 3 and column 2.

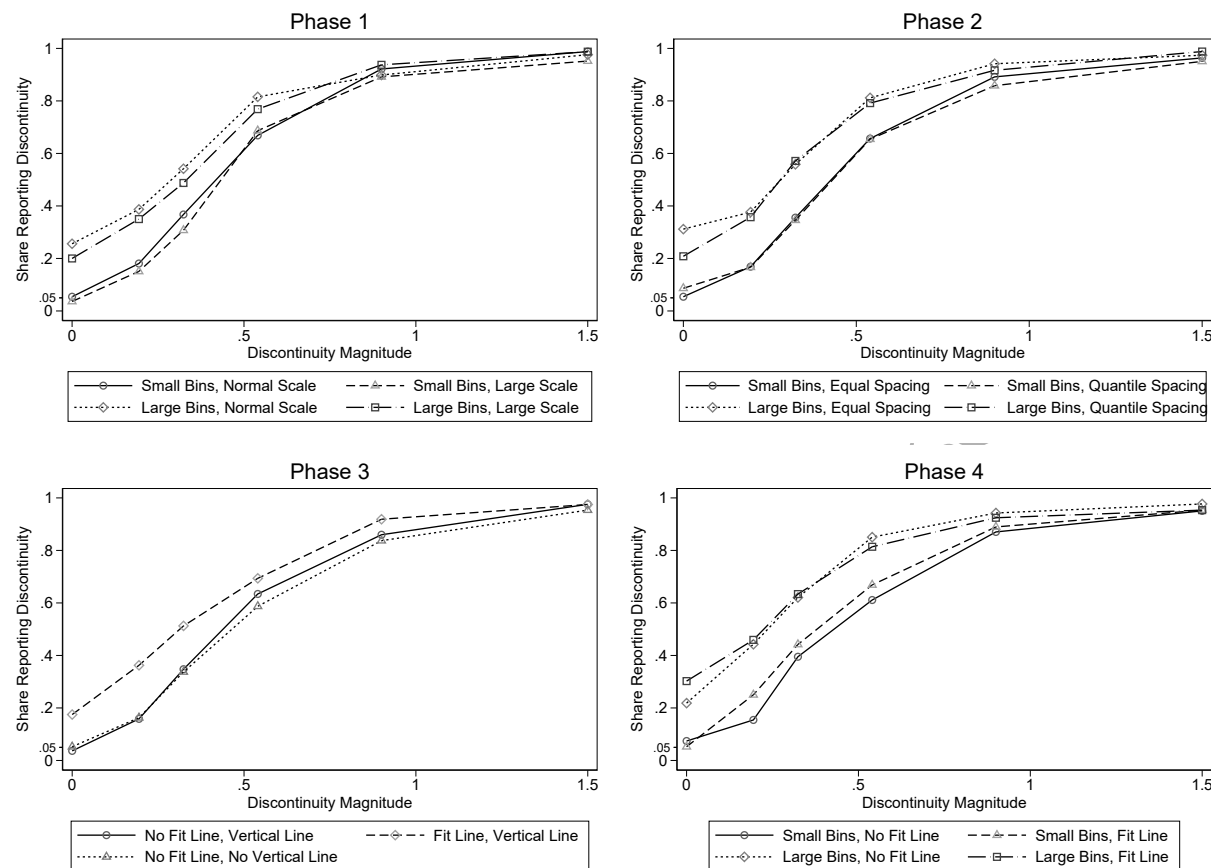


Figure V
Power Functions by Experimental Phase

Notes: Plotted are empirical power functions from the four nonexpert experiment phases. As described in Table I, we test the effects of two graphical parameters in each phase while holding fixed the other three parameters. The legend underneath each graph indicates the combination of graphical parameters each phase tests. 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. *Large bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that minimizes the integrated mean squared error of the bin-average estimators of the conditional expectation function; *Small bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that aims to approximate the variability of the underlying data; *Quantile spacing* indicates that bins were spaced by quantiles rather than evenly spaced; *Fit line* indicates the presence of parametric fit lines; *Vertical line* indicates the presence of a vertical line at the policy threshold; *Normal scale* corresponds to the default y -axis scaling using Stata 14; *Large scale* doubles that default range.

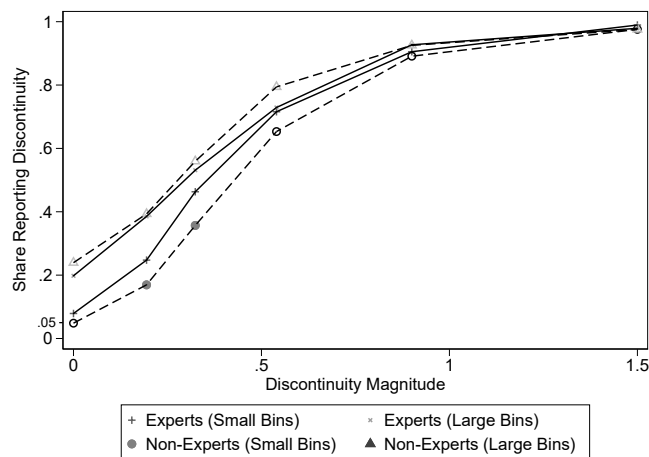


Figure VI
Expert vs Nonexpert Performance

Notes: Plotted are the power functions for the experts and nonexperts. Markers in the figure are shown as solid, matching the legend, whenever nonexperts perform statistically significantly differently at the 5% level from experts under the same graphical treatment and at the same discontinuity magnitude. Markers in the figure are shown as the same shape but hollow instead whenever nonexperts do not perform statistically significantly differently at the 5% level from experts under the same graphical treatment and at the same discontinuity magnitude. *Large bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that minimizes the integrated mean squared error of the bin-average estimators of the conditional expectation function; *Small bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that aims to approximate the variability of the underlying data.

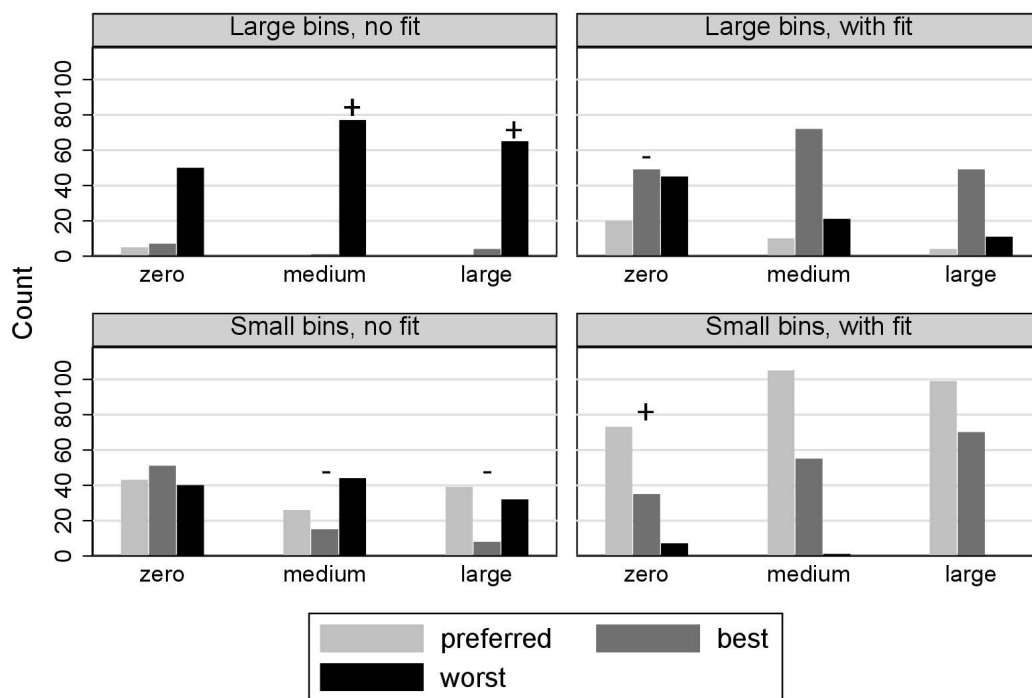


Figure VII
Expert Preferences and Beliefs Regarding Nonexpert Performance

Notes: Each panel 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 nonexpert sample; or believe it to be the worst-performing treatment among our nonexpert sample. For comparison, the treatments that performed best and worst in that sample are marked with a + and – sign respectively. *Large bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that minimizes the integrated mean squared error of the bin-average estimators of the conditional expectation function; *Small bins* corresponds to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that aims to approximate the variability of the underlying data; *with fit/no fit* indicate the presence or absence of parametric fit lines.

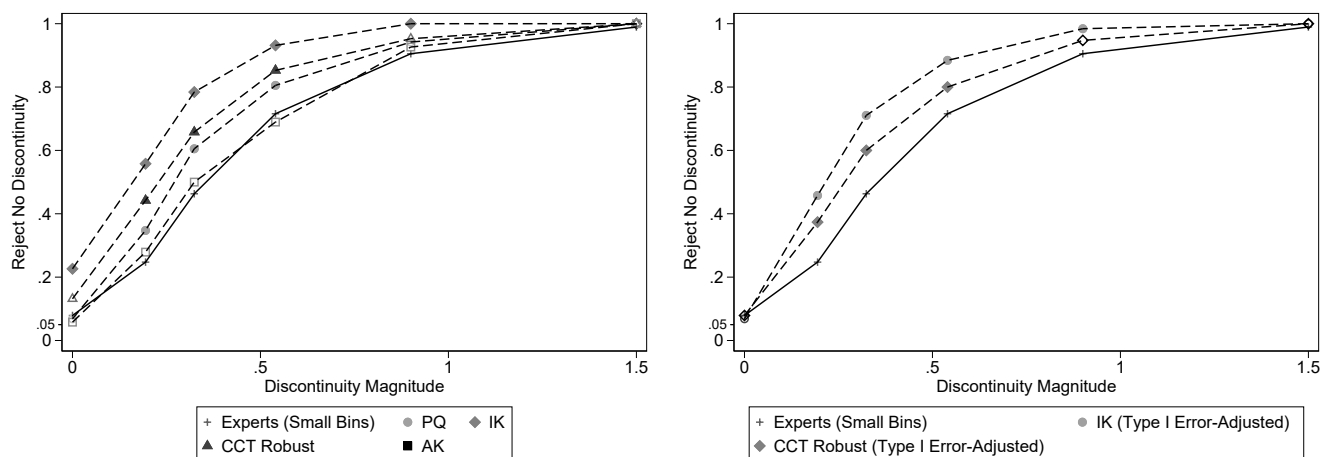


Figure VIII
Expert Visual vs Econometric Inference

Notes: PQ uses a correctly specified regression model with global 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`. AK uses the `RDHonest` procedure with the rule-of-thumb bound on each DGP's second derivative. Markers in the figure are shown as solid, matching the legend, whenever the econometric inference procedure performs statistically significantly differently at the 5% level from expert visual inference at the same discontinuity magnitude. Markers in the figure are shown as the same shape but hollow instead whenever the econometric inference procedure does not perform statistically significantly differently at the 5% level from expert visual inference at the same discontinuity magnitude. Expert visual inference is shown for the case of *small bins*, corresponding to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that aims to approximate the variability of the underlying data.

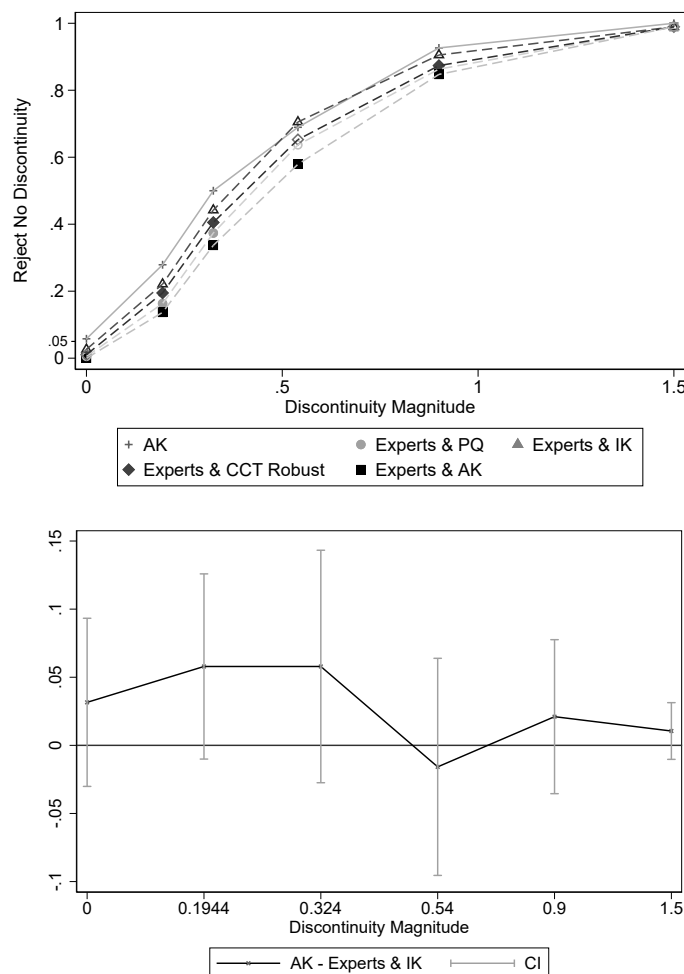


Figure IX
Combined Expert Visual and Econometric Inference vs AK

Notes: Combined expert and visual econometric inference is based on the performance of an inference procedure which infers a discontinuity if and only if both the visual and econometric procedures reject no discontinuity. The relevant expert inference is based on *small bins*, corresponding to the Calonico, Cattaneo, and Titiunik (2015) bin width selector that aims to approximate the variability of the underlying data. PQ uses a correctly specified regression model with global 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`. AK uses the `RDHonest` procedure with the rule-of-thumb bound on each DGP's second derivative. Markers in the figure are shown as solid, matching the legend, whenever the combined econometric-expert-visual inference procedure performs statistically significantly differently at the 5% level from the AK procedure at the same discontinuity magnitude. Markers in the figure are shown as the same shape but hollow instead whenever the combined econometric-expert-visual inference procedure does not perform statistically significantly differently at the 5% level from the AK procedure at the same discontinuity magnitude. In the graph on the bottom, we plot the difference between the combined IK-expert-visual and AK inference procedures, along with 95% two-way cluster-robust confidence intervals per Cameron, Gelbach, and Miller (2011) that account for potential correlation between visual and econometric inferences at the dataset level (see Online Appendix H.1 for details).