Visual Inference and Graphical Representation in Regression Discontinuity Designs

Christina Korting, Carl Lieberman, Jordan Matsudaira, Zhuan Pei, Yi Shen

Presented by Javid Karimli

Dec 08, 2023

https://storage-googleapis-com.ezproxy1.hw.ac.uk/rd-video-turial/rd_video_tutorial.mp4

Motivation

- “Few would deny that the most powerful statistical tool is graph paper.”
    - —Geoffrey S. Watson (1964)
- Effective use of graphs conveys a large set of statistical information at once and improves research transparency (Andrews, Gentzkow, and Shapiro 2020).
- To understand the best use of graphical evidence, it is important to study readers’ ability to process information from graphs—“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.

Motivation

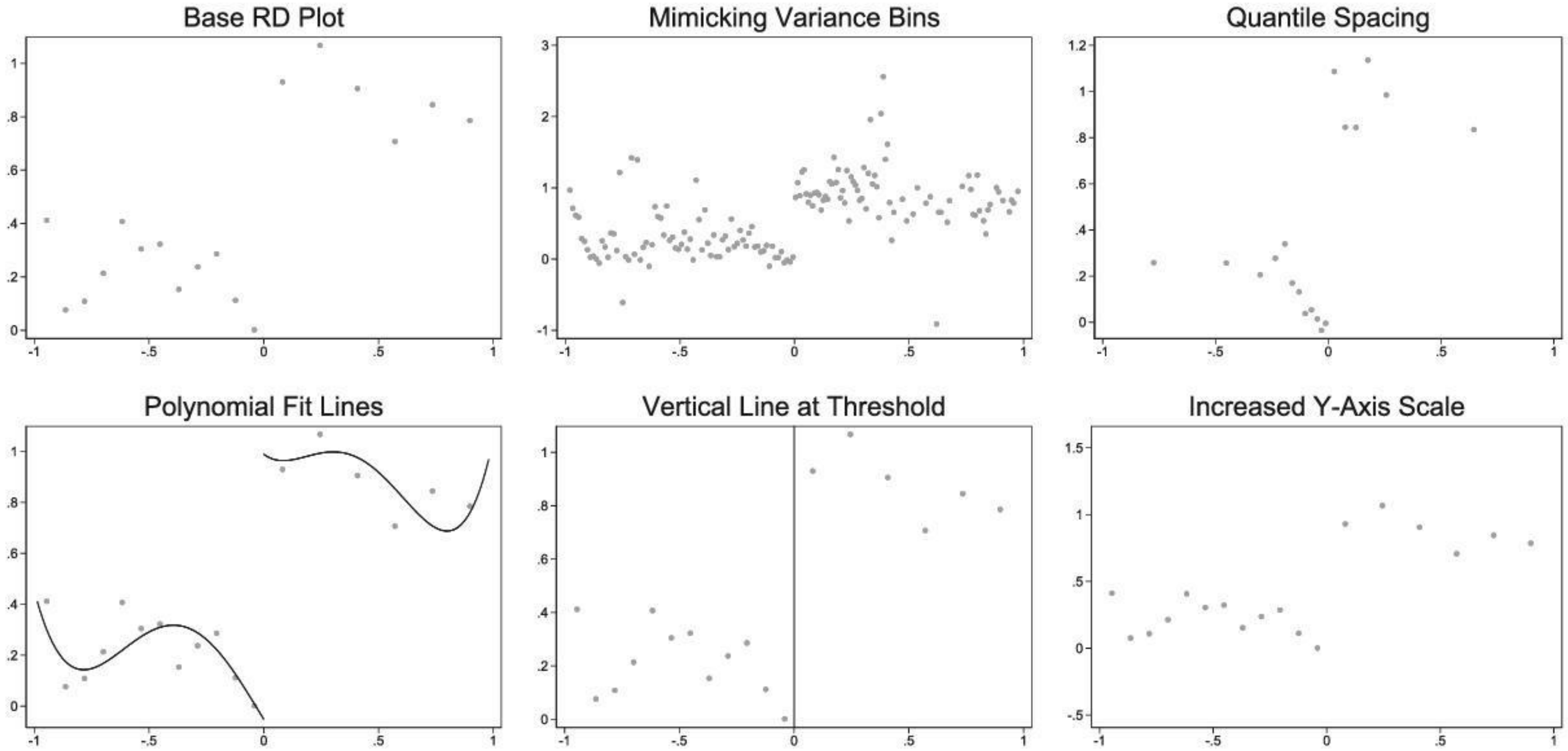
- There is limited research on how to choose graphical parameters in practice.
- Calonico, Cattaneo, and Titiunik (2015) propose two popular data-driven bin width selectors:
  - one that minimizes the integrated mean squared error (IMSE) of the bin averages, resulting in **fewer, larger bins**,
  - and another that mimics the variability of the underlying data (mimicking variance, MV), which leads to **more, smaller bins**.

Notation

- Vector γ denotes a combination of graphical parameters.
    - Five parameters in this article (bin width, bin spacing, axis scaling, polynomial fit lines, and a vertical line at the policy threshold), and each entry of γ represents the value of a particular parameter.
- The combination (g, d) denotes the probability model underlying an RD data set. g encompasses four elements:
    - (i) the distribution of the running variable X;
    - (ii) the conditional expectation function E[Y˜ |X = x], which is continuous at the policy threshold x = 0;
    - (iii) the distribution of the error term u, where Y˜ ≡ E[Y˜ |X = x] + u;
    - (iv) the sample size N.
- Intuitively, g specifies everything in the probability model except for the discontinuity (d), including the shape of the CEF

Figure I Illustration of Graphical Parameters Tested

Plots are based on the original data from DGP9. The “Base RD ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Conceptual Framework

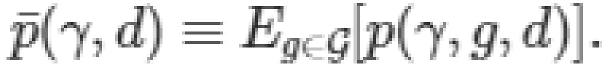
- This is the probability that a randomly chosen reader reports a discontinuity in a graph randomly generated from the GGP (γ, g, d).
- A high value of p indicates a high classification error probability when the true discontinuity d is zero (type I error), but a low classification error probability 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:

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

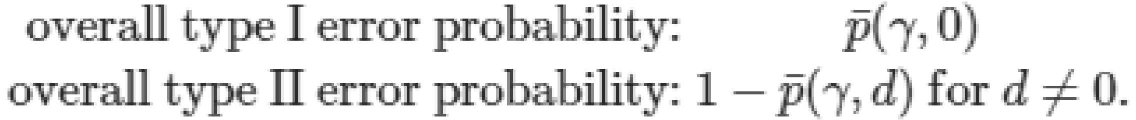
$$\begin{aligned} &g\text{-specific type I error probability:} && p(\gamma, g, 0) \\ &g\text{-specific type II error probability:} && 1 - p(\gamma, g, d) \text{ for } d \neq 0. \end{aligned}$$

Conceptual Framework

- Conceptually, we can further average p(γ, g, d) over the space of DGPs, to arrive at the overall discontinuity classification probability for γ
- Consistent with the definitions in Casella and Berger (2002, 382), we call p(γ, g, d) and $\bar{p}$ (γ, d) power functions as functions of d.
- 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 Ti(γ, g, d), where i takes on values in the set {1, ..., M} and is asked to assess the presence of a discontinuity.



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



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

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

Conceptual Framework

- Let the binary variable Ri(Ti(γ , g, d)) denote participant i’s discontinuity classification, which equals one if the participant reports a discontinuity at the policy threshold
- For a given GGP (γ, g, d), the Ri(Ti(γ, g, d))’s are i.i.d. with E[Ri(Ti(γ, g, d))] = p(γ, g, d).
- A natural estimator for p(γ, g, d) is the sample average of discontinuity classifications:

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

- Specify J DGPs that approximate data from existing research and present graphs generated with discontinuity d for each DGP gj (j = 1, ..., J) to a distinct group of M participants for a total of M · J participants and visual discontinuity classifications.
- The DGP gj’s are randomly sampled from G. A natural estimator for $\bar{p}$(γ, d) is:

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

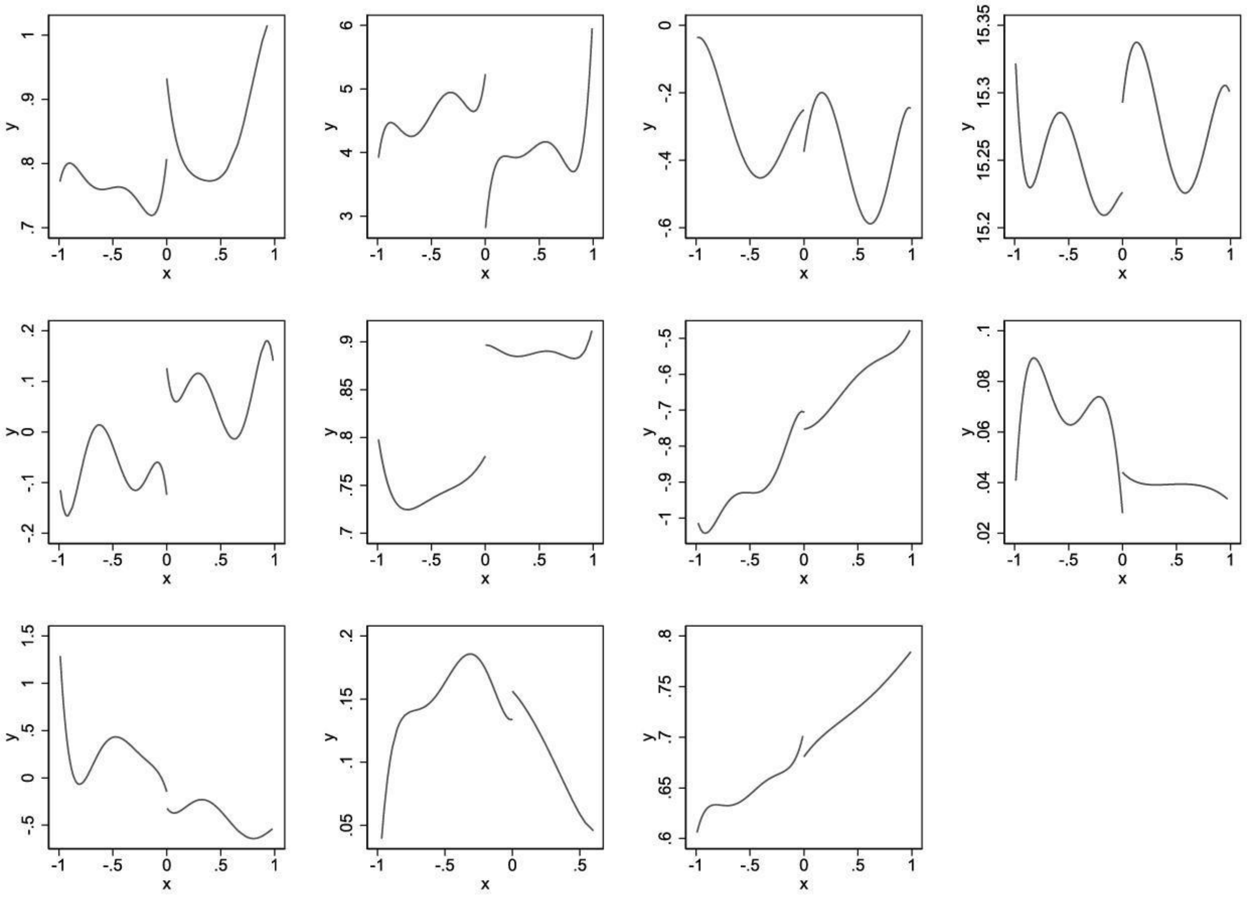
Conceptual Framework

- In summary, we have defined the type I and type II error rates for visual inference:
  - The <u>type I error</u> rate is the fraction of continuous graphs which participants incorrectly classify as having a discontinuity,
  - The <u>type II</u> 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 that a randomly chosen person commits when classifying a graph generated from a representative DGP. These probabilities also help inform best graphical practices.

Experiment Design

- We specify DGPs 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. (DGP1–DGP11).

Figure II Conditional Expectation Functions DGP1–DGP11
Each panel represents the plot of the conditional expectation ...



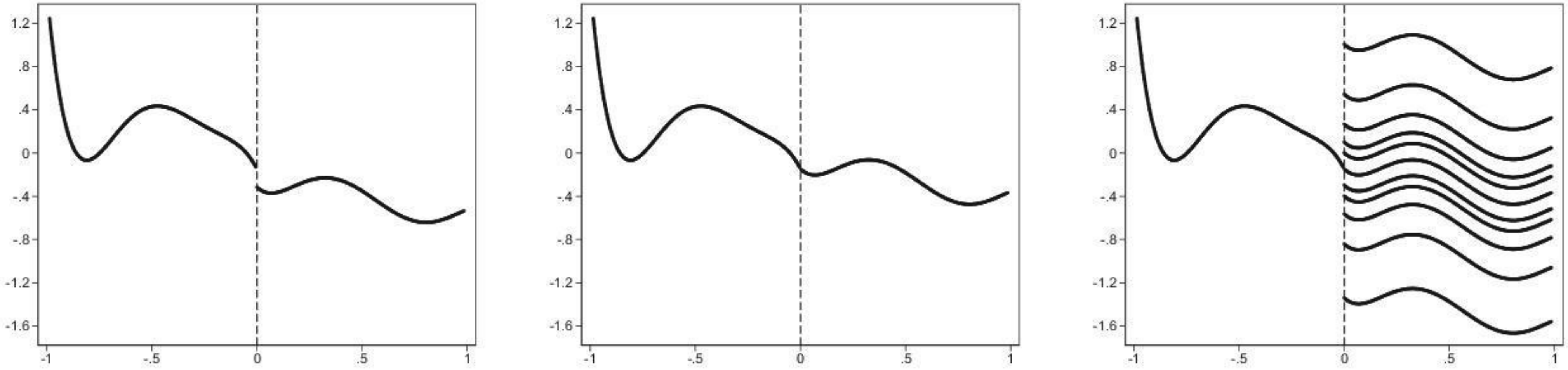
Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Creation of Simulated Datasets and Graphs

- 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.
- As a multiple of σ (error term), d takes on 11 values: 0, ± 0.1944σ, ± 0.324σ, ± 0.54σ, ± 0.9σ, ± 1.5σ. We choose the upper bound |d| = 1.5σ based on our own visual judgment: it represents the point at which we expect every reasonable person to say a graph from any of the 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.

Figure III Creation of Conditional Expectation Functions, DGP9
The leftmost figure plots the piecewise quintic CEF ...



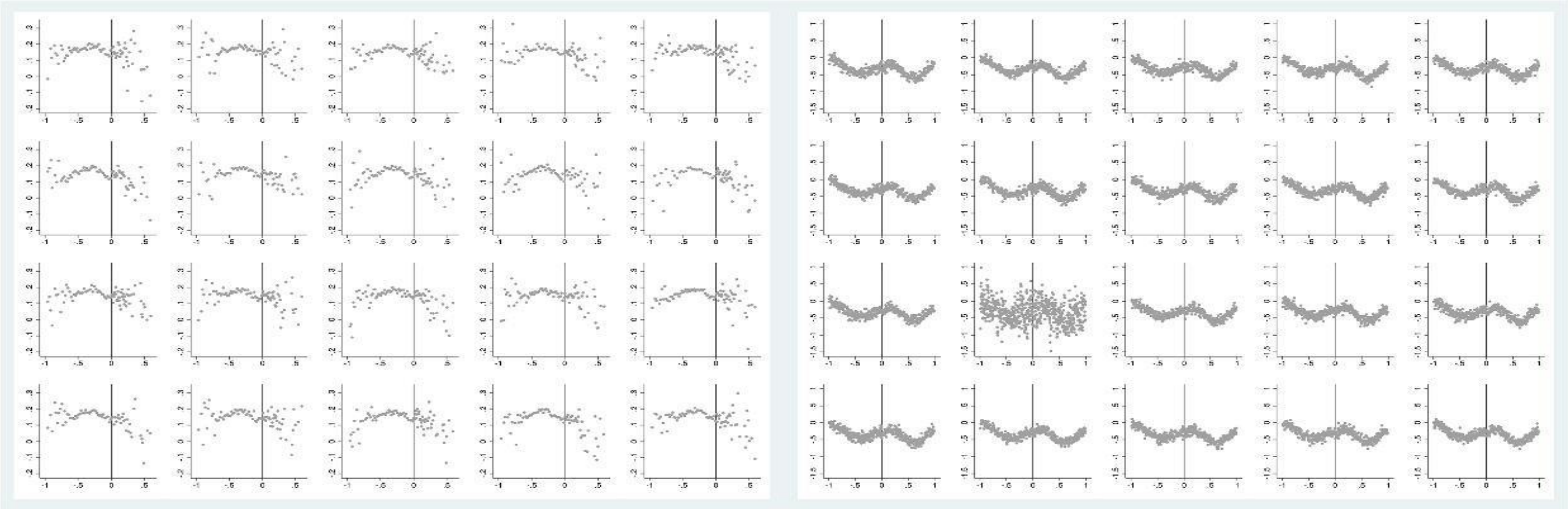
Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Creation of Simulated Datasets and Graphs

- As argued in Section II, we want our DGPs to be representative. Although we select the articles 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 data sets come from the calibrated DGPs. Specifically, we present one graph of the original data randomly placed among 19 graphs from data sets drawn from the corresponding DGP. The goal is to identify the true data set 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 data sets is 5% for a single reader. For our lineup protocol, each graph is a binned scatter plot using the MV bin selector.

Figure IV Lineup Protocol Graph Examples: DGP10 and DGP3
One of the 20 graphs for each lineup protocol is produced from ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Non-expert Experiment Design

- 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; Rosen 2015; Sides 2015), suggesting the participants in our sample should be capable of interpreting the graphs.

Non-expert Experiment Design

- Before the experiment, participants watch a video tutorial explaining how the graphs are constructed. We do not instruct participants on how to make their decisions, for example, whether only to look at points near the cut-off 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 graphs they classify may feature a discontinuity.
- https://storage-googleapis-com.ezproxy1.hw.ac.uk/rd-video-turial/rd_video_tutorial.mp4

Non-expert Experiment Design

- In each phase of the experiment, we present participants with a series of RD graphs using data generated as described before.
- Participants see two graphs with zero discontinuities, one each of ± 0.1944σ, ± 0.324σ, ± 0.54σ, ± 0.9σ, 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.

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.

Non-expert Experiment Design

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

Expert Experiment Design

- 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.

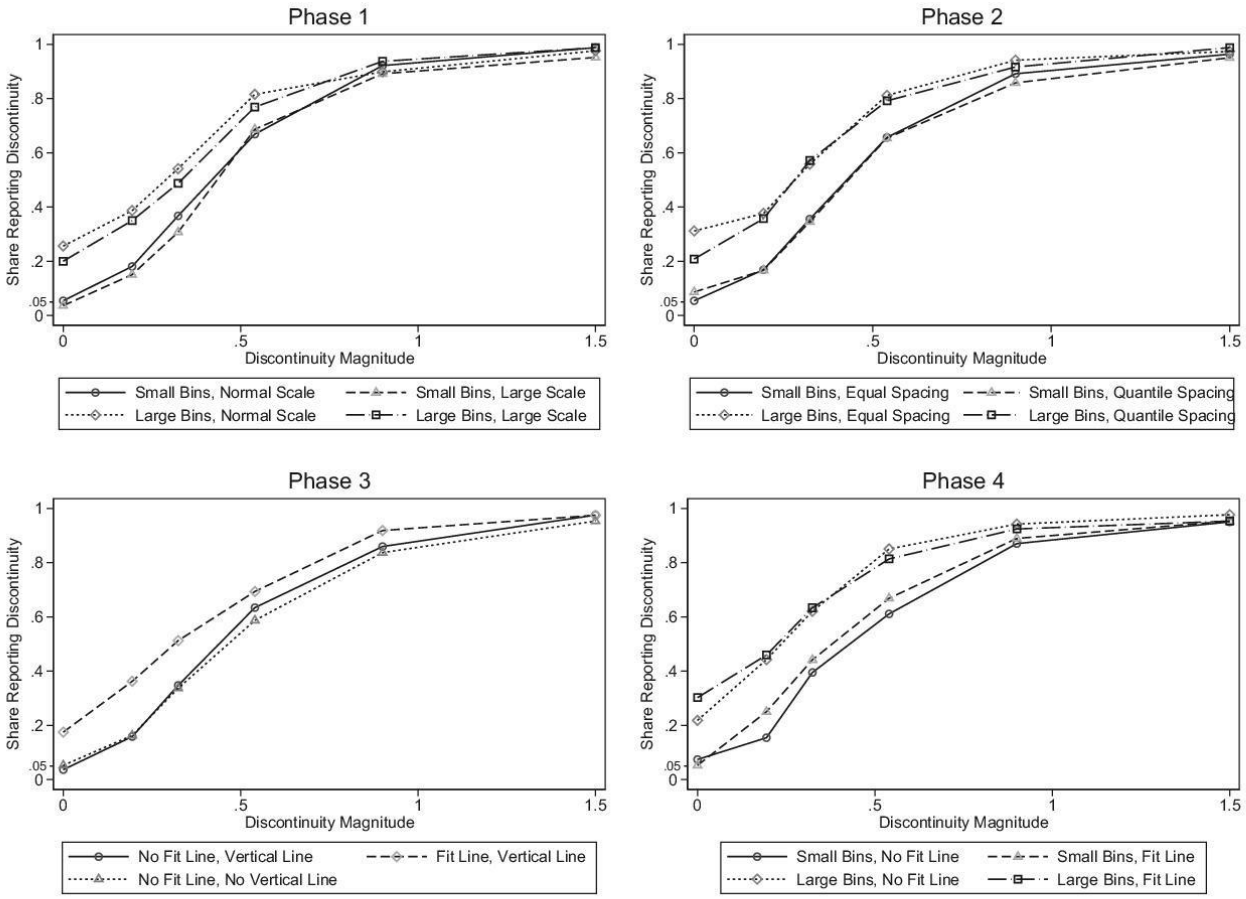
Expert Experiment Design

- 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 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 evenly 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.

Expert Experiment Design

- The second part of the expert study asks about experts’ preferences and their beliefs regarding nonexpert 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.

Figure V Power Functions by Experimental Phase
Plotted are empirical power functions from the four nonexpert ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Results from Non-Expert Experiments

- 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).
  - Large bins have a significantly higher type I error rate relative to small bins
  - The large bins have a type II error advantage over the small bins
  - Axis scaling has little effect on participant perception
- 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.
  - We conclude that bin spacing has a small or null effect on visual inference for the DGPs we test.

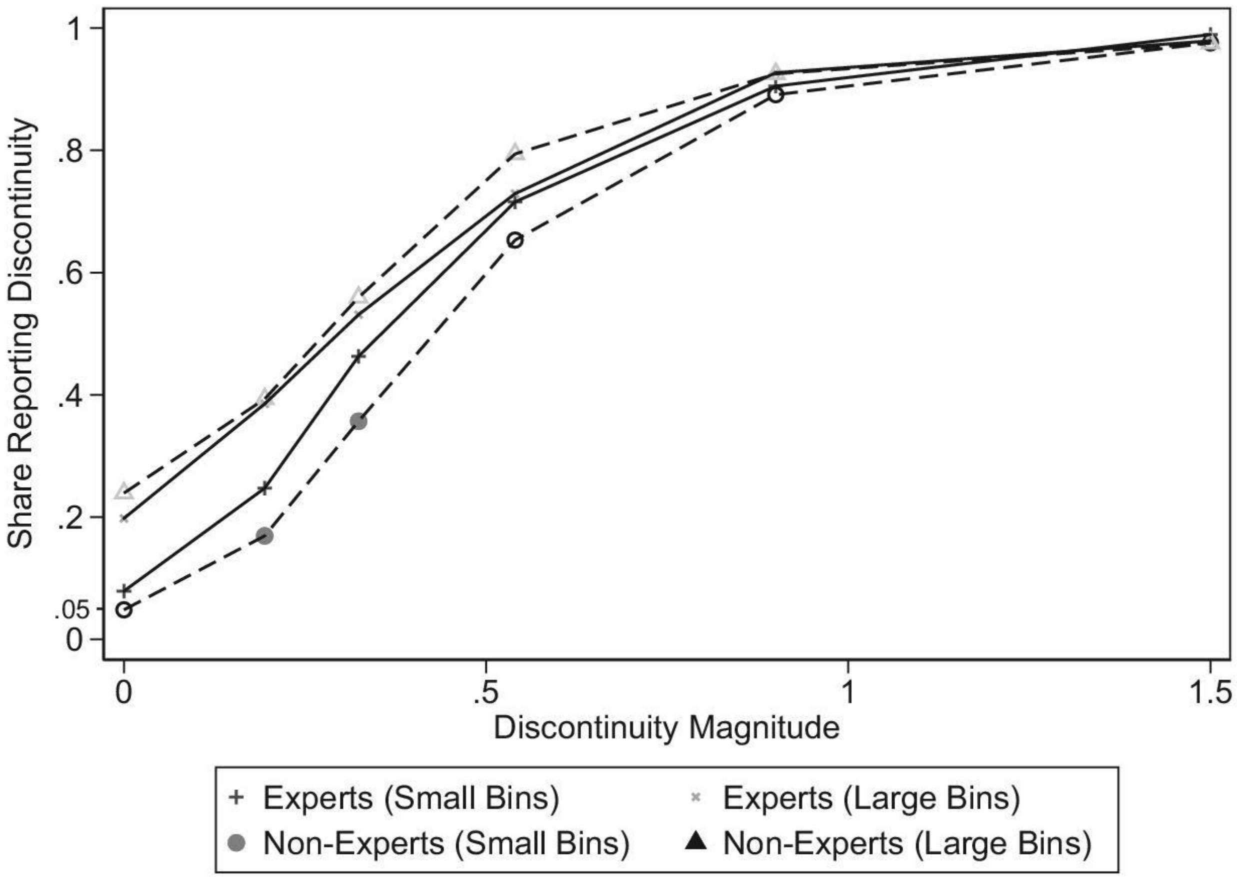
Results from Non-Expert Experiments

- 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 cut-off 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 effect on visual perceptions of discontinuities.

Results from Non-Expert Experiments

- We therefore base our analysis of expert preferences and expert predictions about nonexpert performance on the interaction of two treatments - presence of fit lines and the bin width choice - and run a final phase of experiments, phase 4, directly comparing the four possible treatment combinations.
    - The effects of bin width choice are once again robust across phases, the effects of fit lines are more muted in this phase.
    - This finding suggests that we cannot conclude that fit lines unequivocally result in an increase in type I errors, but they do add uncertainty to visual inference.

Figure VI Expert versus Nonexpert Performance
Plotted are the power functions for the experts and nonexperts.
Markers ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

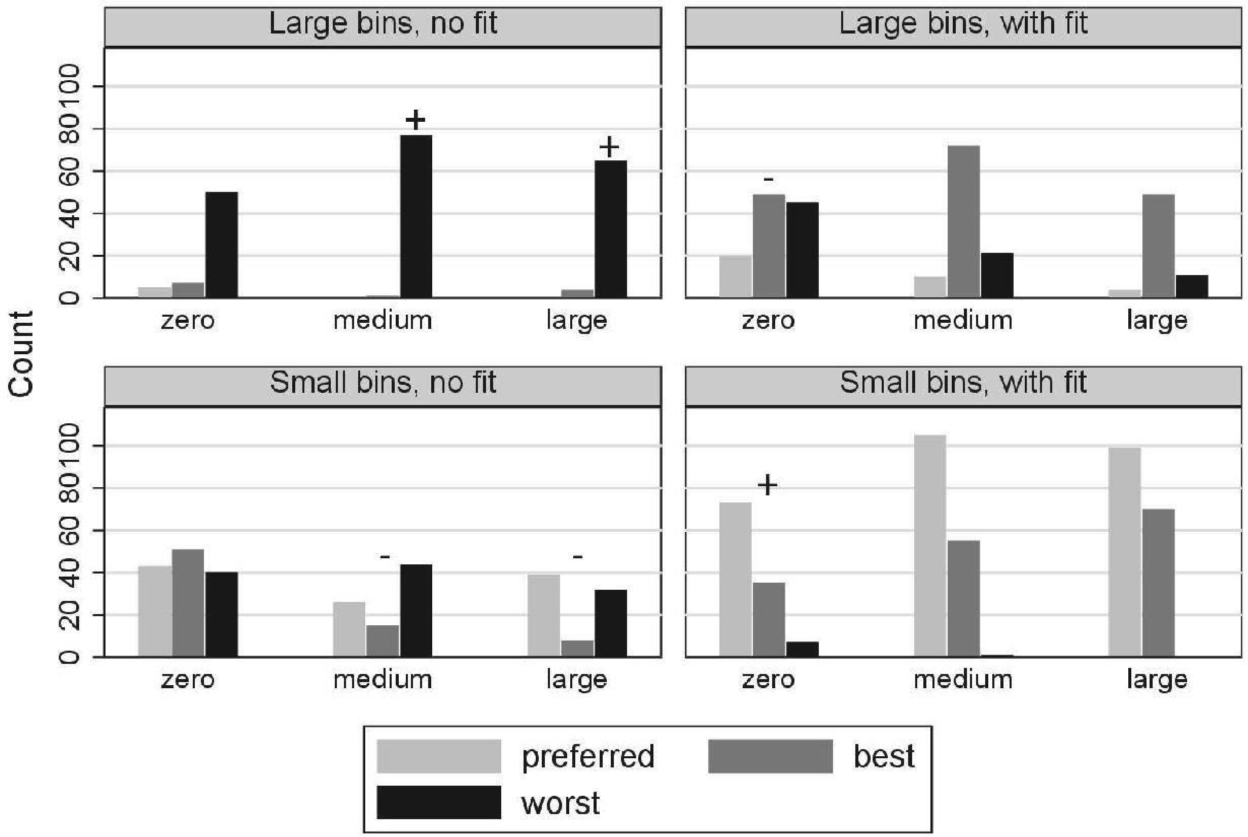
The content of this slide may be subject to copyright: please see the slide notes for details.

Results from Expert Experiment

- We show most expert participants (95 out of 143) graphs generated with our preferred method.
- 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.
- 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.

Figure VII Expert Preferences and Beliefs Regarding Nonexpert Performance

Each panel shows the number of experts who ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

Results from Expert Experiment

- When asked about graphing options for the main graph of an article that conveys the treatment effect, most experts report preferring small bins, usually with fit lines.
- 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

Results from Expert Experiment

- 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.
- 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 versus type II error trade-off 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.

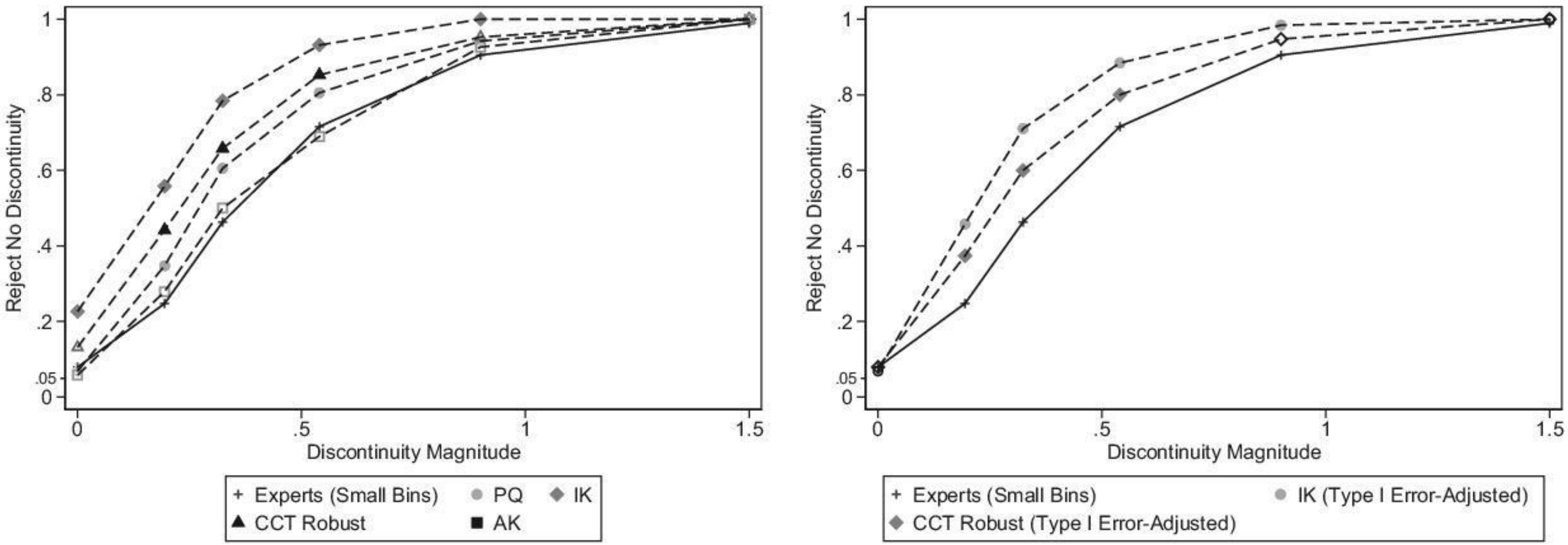
Conclusion/Recommendation

- Based on our experimental results, we recommend the graphical method that uses following features as a sensible default for generating RD graphs.
  - **small bins,**
  - **no fit lines,**
  - even spacing,
  - default y-axis scaling,
  - and a vertical line at the policy threshold
- 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.

Thank You For Your Attention

Attention Check: What colour is the bird of interest?

Figure VIII Expert Visual versus Econometric Inference
PQ uses a correctly specified regression model with global ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

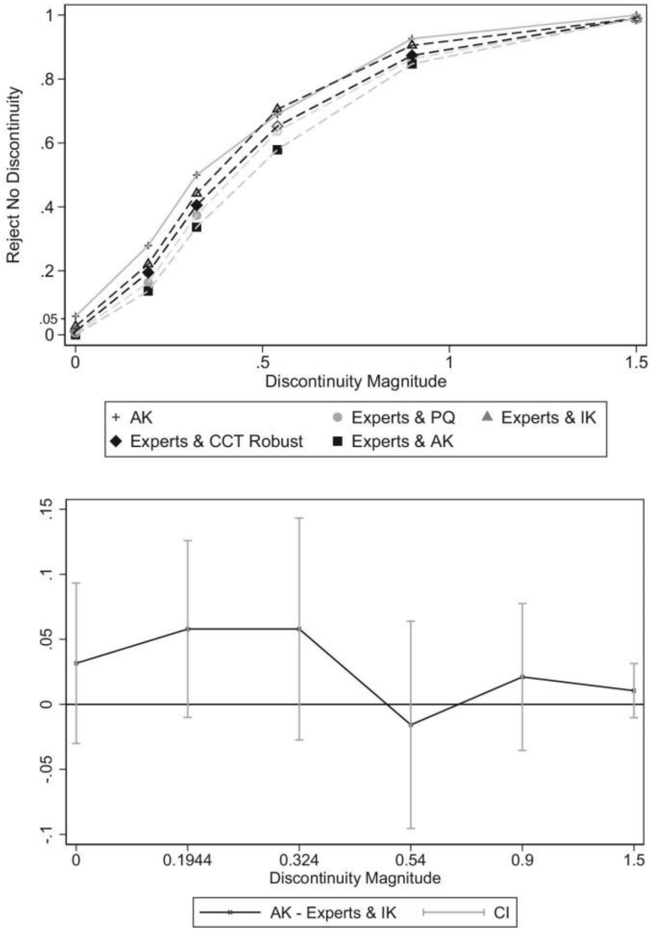
The content of this slide may be subject to copyright: please see the slide notes for details.

Visual Inference vs Econometric Inference

- 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 Kolesar (2018).

Figure IX Combined Expert Visual and Econometric Inference versus AK

Combined expert and visual econometric inference ...



Q J Econ, Volume 138, Issue 3, August 2023, Pages 1977–2019, https://doi.org/10.1093/qje/qjad011

The content of this slide may be subject to copyright: please see the slide notes for details.

OXFORD UNIVERSITY PRESS

Visual Inference vs Econometric Inference

- Furthermore, visual and econometric inferences appear to be complementary.
- Through the analysis of the joint distribution of visual and econometric tests we find that although they commit similar type II errors, there does not appear to be a strong association in their type I errors.