Causal quartets: Different ways to attain the same average treatment effect*

Andrew Gelman, Jessica Hullman, and Lauren Kennedy

24 Sep 2023

Abstract

The average causal effect can often be best understood in the context of its variation. We demonstrate with two sets of four graphs, all of which represent the same average effect but with much different patterns of heterogeneity. As with the famous correlation quartet of Anscombe (1973), these graphs dramatize the way in which real-world variation can be more complex than simple numerical summaries. The graphs also give insight into why the average effect is often much smaller than anticipated.

1. Given that real-world effects vary, and statistics is the study of variation, why does the causal inference literature focus on average effects?

Causal inference in statistics and economics focuses on the average causal effect. The purpose of this paper is to raise awareness of different patterns of heterogeneous causal effects: examples where the average effect does not tell the whole story.

Given that real-world effects vary, and statistics is the study of variation, it seems obvious to look at the variation of causal effects across different populations, different scenarios, different time frames, etc. Indeed, the very phrase "average causal effect" (which can be defined in many different ways (see, for example, Angrist and Pischke, 2009) implicitly considers how the effect might vary; otherwise one could simply say "causal effect" without the modifier.

We are far from alone in advocating for the value of seeking to understand sources of variation. For example, Heckman and Smith (1995) and Heckman, Smith, and Clements (1997) consider the relevance of varying treatment effects in policy analysis; Shadish, Cook, and Campbell (2002) discuss heterogeneity in causal inference; recent authors such as Baribault et al. (2018), Bryan, Tipton, and Yeager (2021), and Yarkoni (2022) have argued for the importance of varying effects, both for theoretical understanding and practical decision making; and Buhl-Wiggers et al. (2023) use Fréchet-Höffding bounds to quantify the implications of varying treatment effects.

Perhaps surprisingly, though, much of the literature of statistics and econometrics focuses on the estimation of average causal effects without much discussion of variation. Before proceeding to discuss the importance of varying treatment effects, it behooves us to consider why there has been such an interest in averages.

There are several good reasons for the traditional approach of considering the treatment effect to be a single parameter to be estimated:

- In a randomized experiment, the average difference comparing treatment and control groups yields an unbiased estimate of the sample average treatment effect (and, if the participants

*We thank Dan Goldstein, Stephen Stigler, Howard Wainer, and two anonymous reviewers for helpful comments and the U.S. Office of Naval Research and Institute of Education Sciences for partial support of this work.

are themselves a random sample of some population, an unbiased estimate of the population average treatment effect). It makes sense to study this average effect, as this is what can be estimated from the data.

- More generally, under different assumptions in observational studies, various average treatment effects are what can be identified (Imbens and Angrist, 1994).

- Under the assumption that the effect does not vary, it can be estimated using a linear regression without interactions, with the coefficient of the treatment variable represents the causal effect. Alternatively, if the underlying effect varies, this coefficient represents an average treatment effect, in the same way that fitting a linear model to nonlinear data can be considered to estimate some sort of average regression line. Hence it can make sense to speak of "the" causal effect in the same way that we would speak of "the" regression coefficient β, as representing a single parameter in a model or a population average quantity. This coefficient is a coherent but perhaps somewhat difficult-to-interpret weighted average of treatment effects.

- Interactions can be hard to estimate; indeed, under some reasonable assumptions you need 16 times the sample size to estimate an interaction than to estimate a main effect (Gelman, 2018). Thus it can make sense to fit a model assuming a constant treatment effect even if you think there may be interactions in reality.

- Under the assumption of a constant treatment effect (the "Fisher null hypothesis"), it is possible to obtain exact confidence intervals for randomized experiments.

- Without additional assumptions, we cannot measure unit-specific treatment effects. Furthermore, while randomization gives us an unbiased estimate of the sample average treatment effect, it doesn't identify variation in treatment effects, as this depends on unmeasurable correlation between potential outcomes.

- Varying treatment effects can themselves be difficult to interpret because the variation is observational: if an effect is higher in some settings than in others, this variation cannot itself necessarily be given a causal interpretation.

For all these reasons, along with the convenience of single-number summaries, it has become standard practice either to fit a model assuming a constant treatment effect or to aggregate to obtain an estimated average treatment effect when fitting models in which effects vary; see, for example, Hill (2011) and Wager and Athey (2018). Subgroup analyses are also performed but are often too noisy to be part of the main conclusions.

That all said, we have become convinced through work in many application areas that thinking about varying effects can be essential for understanding causal inference, and consequently for making decisions based on estimates, such as in implementing policies or interventions beyond the lab. In this article we present causal quartets as a graphical tool for helping reform how we think

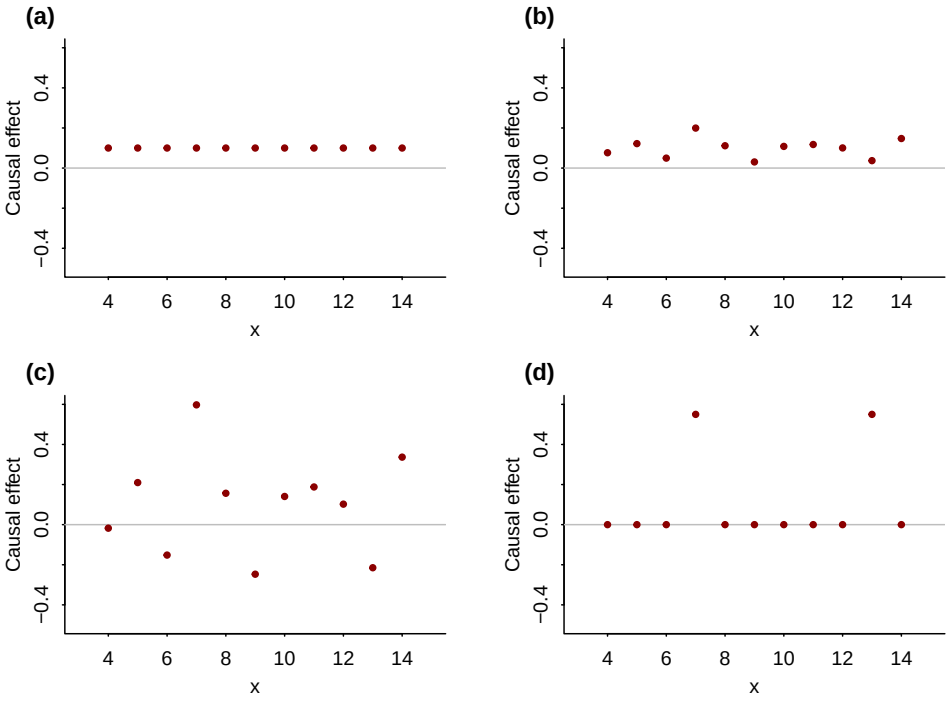

Figure 1: *Four graphs showing different patterns of causal effects, each with average effect of 0.1: (a) constant effect, (b) low variation, (c) high variation, (d) occasional large effects.*

about effects. Section 2 demonstrates and explains the value of such tools. Section 3 presents a software package that researchers or consumers of research can use to create causal quartets in order to reflect on their own research goals or interrogate effects in the literature. In Section 4 we discuss implications of treatment-effect heterogeneity for statistical practice in the context of the reasons discussed above for traditionally focusing on the average.

2. Two causal quartets

2.1. Plots of latent causal effects

We dramatize variation in causal effects with two "quartets": sets of four plots with the same average effect but much different patterns of individual effects. All the displays plot the causal effect vs. a hypothetical individual-level predictor, x. The first quartet shows examples of unpredictable or random variation, so that x is essentially just an index of units. The second quartet shows effects that vary as different systematic functions of x. More generally, x could represent different types of units and could be an observed predictor (such as a pre-test in education), or a latent quantity (such as overall health status in a medical study). The quartets are conceptual plots of different scenarios, not direct graphs of data.

Figure 1 shows four very different scenarios corresponding to an average treatment effect of 0.1. Figure 1a shows the simplest case, often implicitly assumed in discussions of "the" treatment effect. It is hard to think of many effects that are truly constant, so this pattern is more of a baseline than

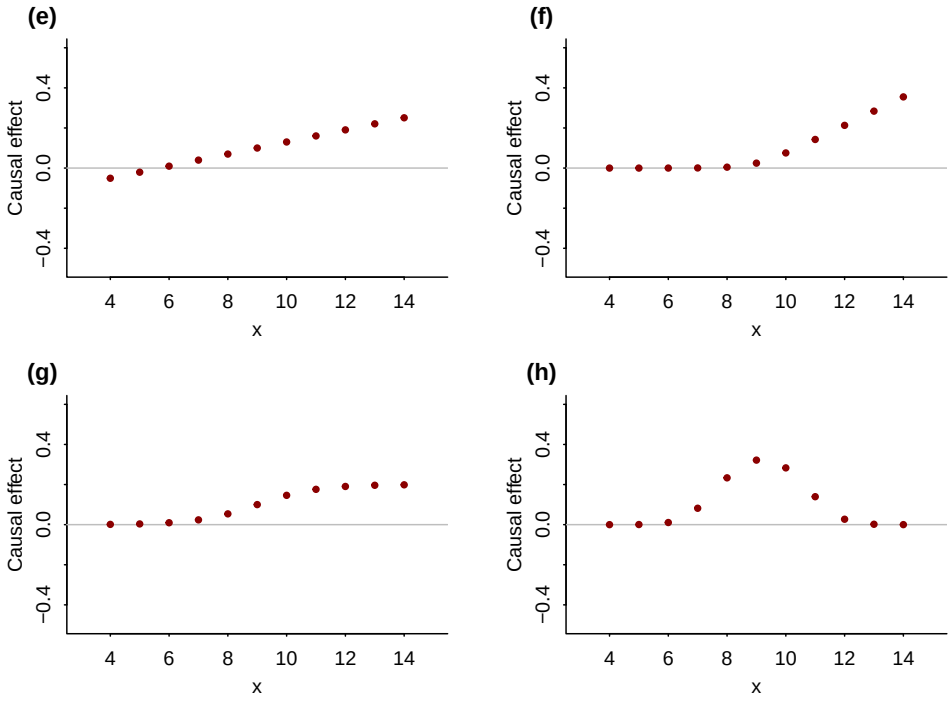

Figure 2: *Four graphs showing patterns of causal effects, each with average effect of 0.1, but varying in different ways as a function of a pre-treatment predictor: (e) linear interaction, (f) no effect then steady increase, (g) plateau, (h) intermediate zone with large effects.*

a realistic choice. Figure 1b shows an effect that is always positive across units but with magnitude varying between 0 and 0.2, for example a medical treatment whose efficacy depends on various unmeasured patient characteristics. In Figure 1c, there is high variation and the effect could be positive or negative at the level of the units; this could be an educational innovation that helps most students but hinders the progress of some students with unusual learning styles. Finally, in Figure 1d the treatment effect is usually zero but is high among some small subset of units with nonzero effects, as could arise with a promotion that is highly effective but only among the small proportion of the target population who are aware of it.

These plots correspond to four different sorts of real-world situations, and we conjecture that some misunderstanding about effect sizes comes from the habit of thinking about the average effect without considering what that means in the context of variation.

Figure 2 presents another quartet, this time showing different forms of interaction in which the effect varies systematically as a function of a pre-treatment predictor. Figure 2e shows a linear interaction, which is typically the first model that is fit when researchers go beyond the model of constant effects. Figure 2f illustrates a treatment that is only effective when the pre-treatment variable exceeds some threshold; this could be a training program that requires some minimum level of preparedness of the trainee. Figure 2g adds a plateau, corresponding to the realistic constraint of some maximum effectiveness. Finally, Figure 2h shows a non-monotonic pattern with a "sweet spot," which could arise in a medical treatment that has no effect for the healthiest patients (because

they do not need the treatment) or for the sickest (for whom the treatment is too late).

The interpretation of Figure 2 will depend on the meaning of the x-variable, which here is shorthand for all the pre-treatment characteristics of the units and conditions of the experiment. The graphs make it clear that changing the range of x can change not only the average treatment effect but also the pattern of interaction. For example, if data arise just from the right half of each graph (that is, $x \geq 10$), then all four treatment effects will look monotonic and indeed not far from linear, which could cause problems with later extrapolation. For that reason, it can be useful when drawing causal quartets to go beyond the data to consider the range of the population of interest, especially recognizing that experiments are often performed on narrow samples and then their conclusions are applied more generally.

As with Figure 1, the second quartet is not intended to represent an exhaustive list of possibilities; Figure 2e shows the typical assumption made in modeling an interaction, while Figures 2f, g, and h represent different sorts of patterns that go beyond what would usually be included in a statistical model. The first quartet shows different levels and distributions of unpredictable variation; the second represents variation that depends on pre-treatment information. A realistic setting would include a mix of both.

Our Figures 1 and 2 are modeled on the famous correlation quartet of Anscombe (1973): four scatterplots where the treatments have the same first and second moments but with much different bivariate patterns. This quartet is useful for teaching the limitations of the correlation statistic and also stimulating students and researchers to consider alternative models for data. Later work has explored general approaches to constructing such plots; see Chatterjee and Firat (2007) and Matejka and Fitzmaurice (2017).

2.2. Plots of observable data

The big difference between our causal quartets and these earlier correlation quartets is that this earlier work concerned plots of *data*, so that departures from the assumed model could be seen directly—hence the title of Anscombe (1973), "Graphs in statistical analysis"—whereas Figures 1 and 2 graph latent *causal effects*, which in general cannot directly be observed. Thus, our plots are conceptual, and their utility to students and researchers is conceptual. Figures 1 and 2 should help in design and analysis of causal studies, both by suggesting ideas for models of treatment effects and as reminders of the limitations of the average causal effect, in the same way that the quartet of Anscombe (1973) dramatized the limitations of the correlation and regression coefficients in descriptive statistics.

To better understand these patterns, it can be helpful to visualize them in terms of observable data. In Figure 3, we display graphs of data that are consistent with the effects shown in Figures 1 and 2. Each of the new plots shows outcomes under treatment and control for the same hypothetical eleven units, with the differences representing the causal effects.

In general, we cannot observe causal effects directly from data: even within-person designs that expose participants to both a control and treatment condition will be affected by factors such as

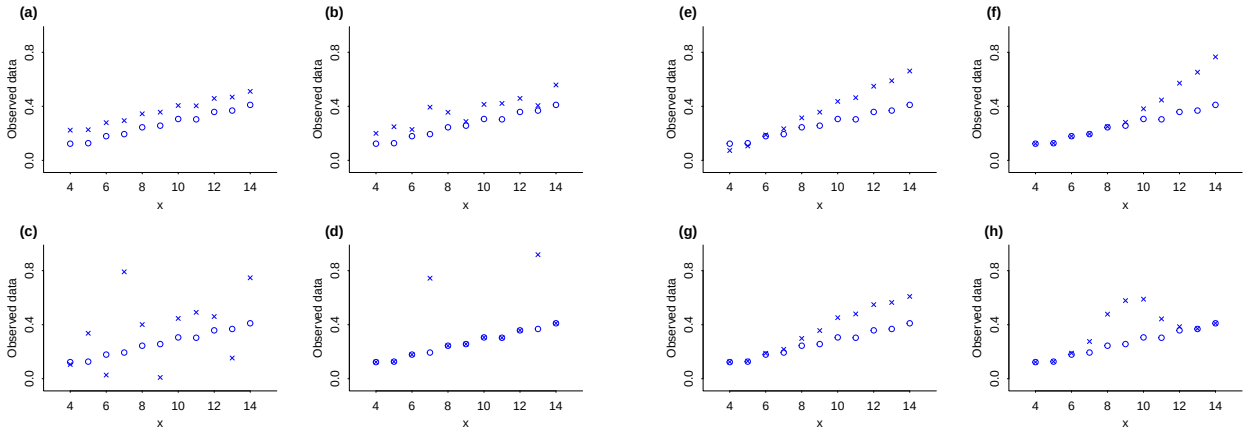

Figure 3: *Two quartets showing different patterns of observable data consistent with the causal effects displayed in Figures 1 and 2. In each plot, the crosses and circles represent treated and control units, respectively, and the difference between the two is the treatment effect.*

order effects. However, the examples in Figure 3 give a sense of what data might look like under different patterns of causal effects in the absence of such factors.

3. causalQuartet: An R package for generating causal quartets

To facilitate generating causal quartets, we created an R package called `causalQuartet`.[1] The package takes as input an average treatment effect and a set of observations x (for latent quartets as in Figures 1 and 2) as well as a set of hypothetical control condition values (for observable quartets as in Figure 3). The user has the option of specifying additional parameters to control the presentation of the quartets.

We envision researchers and consumers of research looking at these quartets of hypothetical treatment effects both before and after a study has been run.

For example, in light of well-known problems with overestimated effects when null hypothesis significance testing is applied to low powered studies (e.g., Button et al., 2013), journalists and other consumers of research might benefit from generating quartets to help explore the implications of a claimed effect size. Researchers who wish to interrogate an effect that they have found or that has been reported by their peers, such as when reviewing a paper, might generate a few quartets to support an argument about why an average treatment effect is likely to be an overestimate.

Before running a study, a researcher can use the quartets in deciding what size effect makes sense to target in sample size calculations. The quartets may also be useful in modeling, where they could help promote thinking about how consistent different patterns would be with prior knowledge. For example, a quartet like Figure 2 could be useful before or after one estimates interactions in a model, to stimulate reflection on the linearity assumption. By aiming to prompt people to reflect on the sorts of informal expectations they bring to data analysis, the package is similar to prior work by Kim, Reinecke, and Hullman (2017) and Hullman et al. (2018), which finds that asking users

[1] `https://github.com/jhullman/causalQuartet`

of graphical displays to make predictions about effects before seeing observed data can improve their recall of the data and their ability to make predictions about new settings. For within-person or pre-post treatment designs, a researcher may even want to compare plots of observed subject-specific differences between a treatment and control to a plot like Figure 3. However, this should be done with acknowledgment that the observables in Figure 3 are hypothetical data that are not subject to factors such as order effects that are generally unavoidable in within-person designs.

Another scenario is one we see far too rarely in intervention-oriented empirical research: reflecting on the utility of putting an intervention into practice given an estimated effect. For example, researchers in disciplines ranging from psychology to medicine to economics to computer science often end their interpretation of estimated effects at the average treatment effect. With the help of causal quartets, researchers can instead use the estimate as a jumping off point for discussing the relative utility to be gained from implementing the new intervention under different assumptions about heterogeneity and varying stakes.

4. Discussion

As has been discussed in the judgment and decision making literature, quantities are generally understood comparatively. Hofman, Goldstein, and Hullman (2020) and Kim, Hofman, and Goldstein (2022) discuss comparisons of effect sizes to inferential or predictive uncertainty. Distributions of effect sizes can be understood in terms of quantiles, which can then map to policy evaluation; see Bitler, Gelbach, and Hoynes (2006). In addition, causal quartet plots can be used to develop intuitions for average treatment effects, which will be relevant in the design and interpretation of studies (Gelman, Hullman, and Kennedy, 2022).

4.1. Different sources of variation in causal effects

Figures 1 and 2 present this potential variation in an abstract way; in particular applications these can represent variation across experimental units, across situations, and over time, and Figure 3 can be used to imagine data consistent with such types of variation, following the UTOS framework of Cronbach (1982). Each type of variation can have applied importance:

- Variation among people is relevant to policy (for example, personalized medicine) and understanding (for example in psychology, as discussed in Gelman, 2014). The quartets can be useful for considering how a policy applied under heterogeneous treatment effects could lead to inequality or fairness; see for example Zidar (2019).

- Variation across situations is relevant when deciding what "flavor" of treatment to do, for example with dosing in pharmacology or treatment levels in traditional agricultural experiments.

- Variation over time is crucial in settings such as A/B testing where an innovation that has been tested on past data is intended to be applied in the future in an evolving business

environment.

Variation in effects is itself important, even setting aside inferential and predictive uncertainty in outcomes, that is, even if the true causal effects are known. That is the point of Figures 1 and 2 and the connection to the quartet of Anscombe (1973): Just as a single number of correlation can represent many sorts of bivariate relationships, so can a single number of average causal effect represent many sorts of causal patterns, even within the simplest setting of a single treatment, a single outcome, and no intermediate variables. One can further distinguish between variation in the effect of a single treatment and variation of treatments (Heiler and Knaus, 2022).

4.2. Why the causal framework?

Nothing in this paper so far requires a causal connection. Instead of talking about heterogeneous treatment effects, we could just as well have referred to variation more generally. Why, then, are we putting this in a causal framework? Why "causal quartets" rather than "heterogeneity quartets"?

Most directly, we have seen the problem of unrecognized heterogeneity come up all the time in causal contexts and not so much elsewhere. We think a key reason is that the individual treatment effect is latent. So it's not possible to make the "quartet" plots with raw data. Instead, it's easy for researchers to simply assume the causal effect is constant, or to not think at all about heterogeneity of causal effects, in a way that's harder to do with observable outcomes. It is the very impossibility of directly drawing the quartets that makes them valuable as conceptual tools.

4.3. Variation and uncertainty

It is said that in the modern big-data world we should embrace variation and accept uncertainty. These two steps go together: modeling of variation is essential for making sense of a world of non-constant treatment effects, but this variation can be difficult to estimate precisely and is sometimes not even identifiable from data, hence the need to accept uncertainty. Just as the quartet of Anscombe (1973) is a reminder of the limits of correlation that is helpful even when our only readily available analytical tool is linear regression, so we hope the quartets in the present paper can help guide us when thinking about generalizing from local causal identification to future prediction and decision making.

References

Angrist, J. D., and Pischke, J. S. (2009). *Mostly Harmless Econometrics.* Princeton University Press.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician* **27**, 17–21.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., ... and Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, **115**, 2607–2612.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review* **96**, 988–1012.

Bryan, C. J., Tipton, E., and Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behavior* **5**, 980–989.

Buhl-Wiggers, J., Kerwin, J., Muñoz, J. S., Smith, J., and Thornton, R. (2023). Some children left behind: Variation in the effects of an educational intervention. *Journal of Econometrics*. `https://doi.org/10.1016/j.jeconom.2021.12.010`

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

Chatterjee, S., and Firat, A. (2007). Generating data with identical statistics but dissimilar graphics: A follow up to the Anscombe dataset. *American Statistician* **61**, 248–254.

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Gelman, A. (2014). When there's a lot of variation, it can be a mistake to make statements about "typical" attitudes. Statistical Modeling, Causal Inference, and Social Science, 8 Oct. `https://statmodeling.stat.columbia.edu/2014/10/08/var/`

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect. Statistical Modeling, Causal Inference, and Social Science, 15 Mar. `https://statmodeling.stat.columbia.edu/2018/03/15/need16/`

Gelman, A., Hullman, J., and Kennedy, L. (2023). Thinking about variation when hypothesizing a plausible average treatment effect. Technical report, Department of Statistics, Columbia University.

Heckman, J., and Smith, J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives* **9**, 85–110.

Heckman, J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies* **64**, 487–535.

Heiler, P., and Knaus, M. (2022). Effect or treatment heterogeneity? Policy evaluation with aggregated and disaggregated treatments. `https://arxiv.org/abs/2110.01427`

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Hofman, J. M., Goldstein, D. G., and Hullman, J. (2020). How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. *Proceedings of the 2020 ACM Conference on Human Factors in Computing Systems (CHI '20)*, 327.

Hullman, J., Kay, M., Kim, Y., Shrestha, S. (2017). Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics* **24**, 446–456.

Imbens, G. W., and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475.

Kim, Y., Hofman, J. M., and Goldstein, D. G. (2022). Putting scientific results in perspective: Improving the communication of standardized effect sizes. *Proceedings of the 2022 ACM Conference on Human Factors in Computing Systems (CHI '22)*, 625.

Kim, Y., Reinecke, K., and Hullman, J. (2017). Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1375–1386.

Matejka, J., and Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, 1290–1294.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences* **45**, E1.

Zidar, O. (2019). Tax cuts for whom? Heterogeneous effects of income tax changes on growth and employment. *Journal of Political Economy* **127**, 1437–1472.