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

Andrew Gelman and Lauren Kennedy

25 Jan 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.

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" implicitly considers how the effect might vary; otherwise one could simply say "causal effect" without the modifier.

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 excellent 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. 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 local average treatment effects are what can be identified (Imbens and Angrist, 1994).

- When estimating a causal model using linear regression without interactions, so the coefficient of the treatment variable represents the causal effect. In the presence of varying treatment effects, 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, so 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.

- 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, 2018a). Thus it can make sense to fit a model with constant treatment effect even if you think there maybe interactions in reality.

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

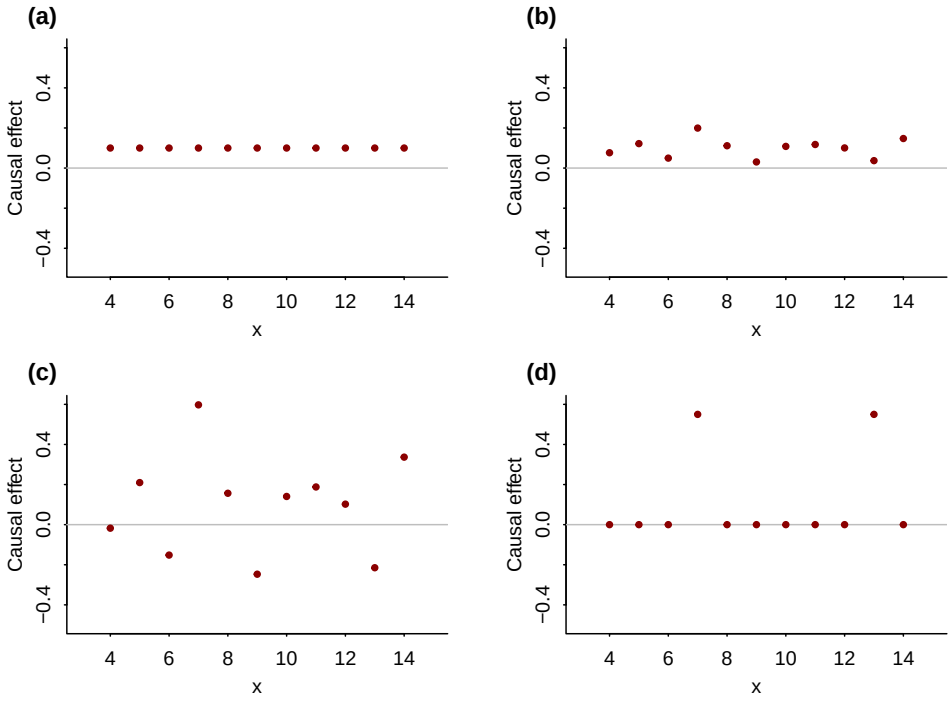

Figure 1: *Four graphs showing different patterns of causal effects, each with average effect of 0.1: (a) constant effect, (b) low variation, (c) high variation, (d) occasional large effects.*

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

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).

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. Sections 2 and 3 of this article explains why, and 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 or a latent quantity. The quartets are conceptual plots of different scenarios, not direct graphs of data.

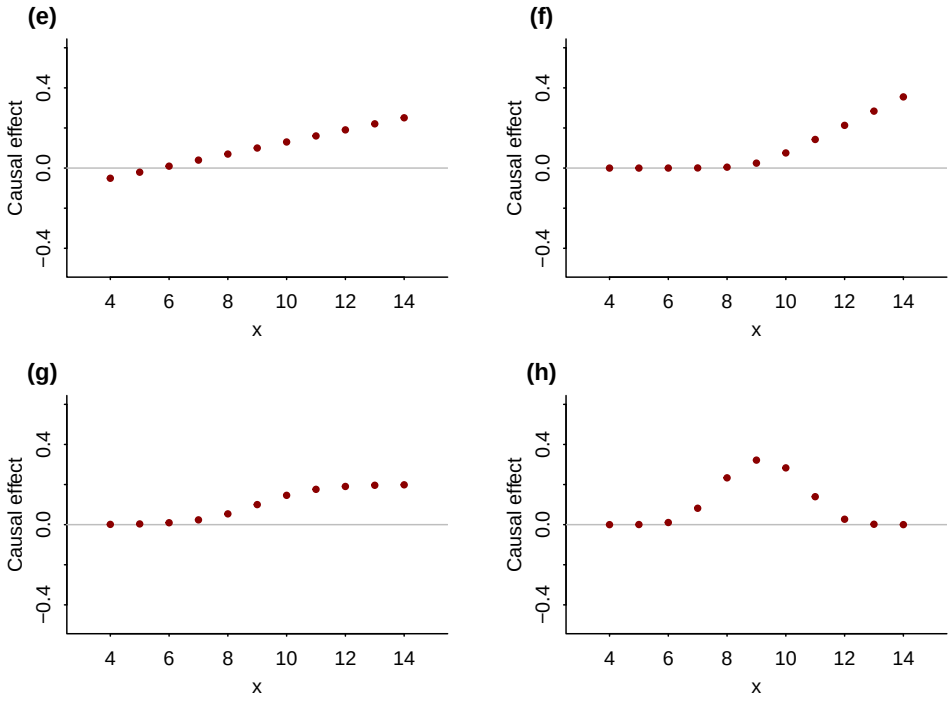

Figure 2: *Four graphs showing patterns of causal effects, each with average effect of 0.1, but varying in different ways as a function of a pre-treatment predictor: (e) linear interaction, (f) no effect then steady increase, (g) plateau, (h) intermediate zone with large effects.*

Figure 1 shows four very different scenarios of corresponding to an average treatment effect of 0.1. Figure 1a shows the simplest case, often implicitly assumed in discussions of "the" treatment effect. Figure 1b shows an effect that is always positive but with magnitude varying between 0 and 0.2. In Figure 1c, there is high variation and the effect could be positive or negative. Finally, in Figure 1d the treatment effect is usually zero but is high among some small subset of units with nonzero effects.

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. We discuss this in the context of some examples in Section 3.

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).

As with Figure 1, these are not intended to represent an exhaustive list of possibilities; they just 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

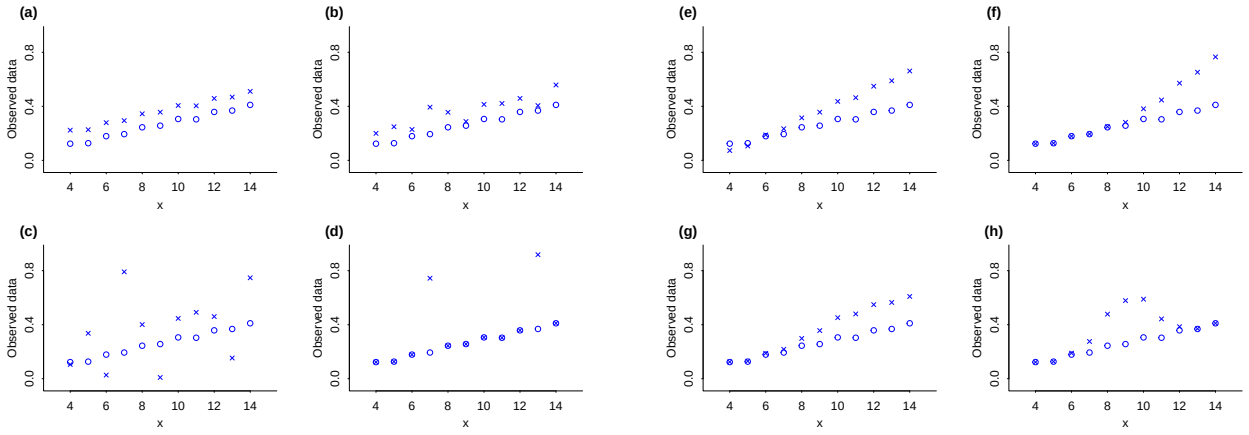

Figure 3: *Two quartets showing different patterns of observable data consistent with the causal effects displayed in Figures 1 and 2. In each plot, the crosses and circles represent treated and control units, respectively, and the difference between the two is the treatment effect.*

include a mix of both.

Our Figures 1 and 2 are modeled on the famous correlation quartet of Anscombe (1973): four scatterplots with 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 we are graphing 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.

Figures 1 and 2 show different patterns of causal effects, all consistent with an average causal effect of 0.1. These are plots of latent data: individual causal effects cannot be directly observed as this would require knowing the value of the outcome under both treatment options.

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 figures shows outcomes under treatment and control for the same hypothetical eleven units, with the differences representing the causal effects. In general, it would not be possible to observe both treatment and control for the same unit, but these graphs give a sense of the sort of data that correspond to different patterns of treatment effects.

3. Some practical implications of the causal quartets

3.1. Anticipating an effect size in a new study

Part of designing a study is accounting for uncertainty in effect sizes. Unfortunately there is a tradition in clinical trials of making optimistic assumptions that allow claims of high power. Here is an example from Zelner et al. (2021). A doctor was designing a trial for an existing drug that he thought could be effective for high-risk coronavirus patients. He contacted us to check his sample size calculation: under the assumption that the drug increased survival rate by 25 percentage points, a sample size of $n = 126$ would assure 80% power. (With 126 people divided evenly in two groups, the standard error of the difference in proportions is bounded above by $\sqrt{0.5 * 0.5/63 + 0.5 * 0.5/63} = 0.089$, so an effect of 0.25 is at least 2.8 standard errors from zero, which is the condition for 80% power for the comparison.) When we asked the doctor how confident he was in his guessed effect size, he replied that he thought the effect on these patients would be higher and that 25 percentage points was a conservative estimate. At the same time, he recognized that the drug might not work. We asked the doctor if he would be interested in increasing his sample size so he could detect a 10 percentage point increase in survival, for example, but he said that this would not be necessary.

It might seem reasonable to suppose that a drug might not be effective but would have a large effect in case of success. But this vision of uncertainty has problems. Suppose, for example, that the survival rate was 30% among the patients who do not receive this new drug and 55% among the treatment group. Then here is a hypothetical scenario of 1000 people:

- 300 would live either way
- 450 would die either way
- 250 would be saved by the treatment

There are other possibilities consistent with a 25 percentage point average benefit—for example the drug could save 350 people while killing 100—but it makes sense to stick with the simple scenario for now. In any case, the point is that the posited benefit of the drug is not a 25 percentage point benefit for each patient; rather, it's a 100% benefit for 25% of the patients.

From that perspective, once we accept the idea that the drug works on some people and not others—or in some comorbidity scenarios and not others—we realize that "the treatment effect" in any given study will depend entirely on the patient mix. There is no underlying number representing the effect of the drug. Ideally one would like to know what sorts of patients the treatment would help, but in a clinical trial it is enough to show that there is some clear average effect. The point is that, once we consider the treatment effect in the context of variation among patients, as in Figure 1c, this can be the first step in a more grounded understanding of effect size.

3.2. Downgrading an apparently huge effect

Gertler et al. (2013) performed a randomized evaluation of an early-childhood intervention program and found it to increase adult earnings by 42%. This sounds a bit too good be true, even more so when considering it as an *average* effect, given that the actual effect must surely vary a lot by person, given the tortuous path from an intervention at age 4 to earnings at age 24. A realistic scenario might be some mix of Figure 1b and d—effects that are often negligible and can follow a wide range when positive—and Figure 1d—an effect that is larger in some intermediate zone. In any of these cases, we would argue that an average effect of 42% is hard to believe, given that it

would reflect some combination of many effects near zero and some increases in earnings of 100% or more.

The implication of this reasoning is that the claimed effect is likely to be a wild overestimate—a point that we earlier made on inferential grounds (Gelman, 2018b) but without reference to varying effects. The framework of nonconstant treatment effects gives us another reason to be skeptical about the claims made for this particular class of interventions.

3.3. Recognizing that an apparently large effect can be explained as an artifact of noise

Beall and Tracy (2013) performed two small surveys and found that women were three times as likely to wear red or pink during certain days of their monthly cycle. Such a large effect is implausible on its face and even more outlandish when considered as an average effect, once we reflect that the effect will be zero for many people, for example, those who never wear red clothing or those whose clothing choices are restricted because of work. Even if there were a factor-of-3 effect for some women in the study, the average effect including those with no effect would have to be much lower, indeed in this case lower than the uncertainty in the estimated effect, which implies that the published result, despite its apparent statistical significance, could be explained by a combination of chance and unintentional selection bias. Indeed, followup studies by these authors and others did not replicate the finding (see, for example, Hone and McCullough, 2020).

Beyond all this, if time of the month does influence clothing choices, we would expect this effect to vary greatly across people and scenarios. Indeed there is no theoretical reason to expect a common direction, hence a pattern such as Figure 1c seems likely, to the extent there are effects at all. Such variation makes it even more difficult to estimate an average treatment effect, as well as implying that any realistic average would be close to zero.

We have used the day-of-cycle and clothing study as an example of the perils of naive interpretation of statistics (Gelman, 2013). Thinking about varying effects helps us understand why estimating an average effect here is not so interesting: the problem is not just the lack of successful replication but rather the conceptual framework under which the effect is characterized by a single number or even a single direction.

3.4. The decline effect: treatments that are less effective in real life than in the literature

When designing a medical trial, the first goal is to maximize statistical power. We say this not cynically but out of a realistic understanding that success—in the form of statistical significance at the conventional level—can be necessary for approval of a drug or procedure, so if you believe your idea is a good one, you want to design your experiment to have the best chance of demonstrating that it works.

Methods of increasing statistical power in an experiment include: (1) increasing the sample size, (2) improving the accuracy of measurements, (3) including additional pretreatment predictors, (4) performing within-person comparisons, and (5) increasing the magnitude of the average treatment effect. Assuming the first four of these steps have been done to the extent possible, one way to achieve the fifth step is to restrict the participants of the study to those for whom the expected effect is as large as possible. To the extent that the treatment effect varies in a predictable way, as in Figure 2, this can be done.

This sort of restriction is perfectly fair, and the result should be a higher average treatment effect among participants in the experiment. When generalizing to a larger population, however, some modeling is necessary conditional on any information used in patient selection. Thinking

about variation in treatment effects makes this clear: the average effect is not a general parameter; it depends on who is being averaged over.

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. In the present paper we compared the average causal effect to its variation.

Figures 1 and 2 present this potential variation in an abstract way; in particular applications these can representation variation across experimental units, across situations, and over time. Each of this sort 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).
- 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.

Even setting aside inferential and predictive uncertainty in outcomes—that is, even if the true causal effects are known—, variation in effects is itself important. 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.

4.1. Why the causal framework?

Nothing in this paper so far requires a casual 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, as in the examples in Section 3, 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, and it's often possible for researchers to assume the causal effect is constant, or to just not think 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.

4.2. The replication crisis

The ideas of this paper have has several points of connection to the replication crisis in science:

- Most immediately, in a world of varying effects, there is no particular interest in testing a null hypothesis of exactly zero effect, and we should be able to move away from the idea that a "statistical significant" finding represents something that should replicate.

- As illustrated in some of the examples in Section 3, when we think about how an effect can vary, we often lower our expectations of its average effect, which in turn can make us aware of problems of low power. For example, if a study is designed under the naive expectation of an effect size of 0.5, but then on reflection we think that an average effect of 0.1 is more plausible, then the study would require 25 times the sample size (or measurements that are 5 times as accurate) in order to maintain the desired power.

- Moving away from the framing of "the" treatment effect helps us think about variation. Instead of classifying a new study as an exact replication (with the implication that the effect should be the same as in the original study) or a conceptual replication (with the hope that the effect should have the same sign), we can think of the first study and the replication as representing two different collections of participants and situations.

As we have argued elsewhere (Gelman, 2015), "once we accept that treatment effects vary, we move away from the goal of establishing a general scientific truth from a small experiment, and we move toward modeling variation (what Rubin, 1989, calls response surfaces in meta-analysis), situation-dependent traits (as discussed in psychology by Mischel, 1968), and dynamic relations (Emirbayer, 1997). We move away from is-it-there-or-is-it-not-there to a more helpful, contextually informed perspective."

For example, consider a hypothetical experiment yielding an estimated treatment effect of 0.003 with standard error 0.001, in a setting in which an effect size of 0.1 would be large. One might first want to dismiss the result as "statistically significant but not practically significant"—but there are various scenarios under which even a small effect would be notable if its sign is well identified. In an A/B testing setting in a large company, even an effect of 0.003 could represent many dollars, and in social science we might be interested in the direction of an effect (for example, knowing whether people under stress performed better or worse on a certain task) more than its magnitude. In such an example, our concern would be that, even if the effect is accurately estimated at 0.003 for this particular experiment, it could easily differ for a new group of people in a different environment. Perhaps the effect would be -0.004 tomorrow, $+0.001$ the next day, and -0.002 the day after that. The relevant comparison is not to the standard error—although that does give us a baseline level of uncertainty—but to changes among people, across scenarios, and over time. Some of this can be learned from data, other aspects of this variation need to be assumed—but there is generally no good reason to assume that the variation in the treatment effect is zero.

A slightly different argument is that in some applications we really only care about the existence and sign of an effect, not its magnitude: knowing that an intervention works, even a small amount, could give insight and be relevant for future developments. But the same problem arises here as before: there is not necessarily any good reason to believe that a small positive effect in one study will apply elsewhere. It is not clear how to interpret an average treatment effect, even in a clean randomized experiment, without considering how the effect could vary across people and scenarios and over time.

4.3. Recommendations for design and analysis

Looking forward, how should this affect statistical practice?

To start, with smaller average effect sizes than previously imagined, better designs are needed: more accurate measurements, better pre-treatment predictors, larger sample size, and within-person comparisons.

When moving to analysis, interactions are important but hard to estimate with precision. So when we do include interactions in our model, we should estimate them using regularization and not

demand that they attain statistical significance or any other threshold representing near-certainty.

Conversely, when we fit simple models without interactions, we should not expect that the local average treatment effects being estimated to immediately generalize. Instead, when generalizing we should allow for both predictable and unpredictable variation in effects, even if in doing so we need to hypothesize scales of variation without direct evidence from the data at hand.

When generalizing beyond the observed sample, it is important to account for changes, which can be done by fitting a model accounting for key pre-treatment variables and then poststratifying to estimate the average treatment effect in the new setting (Kennedy and Gelman, 2021).

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

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

Beall, A. T., and Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological Science* **24**, 1837–1841.

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.

Dehejia, R., and Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology* **103**, 281–317.

Gelman, A. (2013). Too good to be true. *Slate*, 24 July. `https://slate.com/technology/2013/07/statistics-and-psychology-multiple-comparisons-give-spurious-results.html`

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/theres-lot-variation-can-mistake-make-statements-typical-attitudes/`

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* **41**, 632–643.

Gelman, A. (2018a). 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/need-16-times-sample-size-estimate-interaction-estimate-main-effect/`

Gelman, A. (2018b). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, 16–23.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeerch, C., Walker, S., Chang, S. M., and Grantham-McGregor, S. (2013). Labor market returns to early childhood stimulation intervention in Jamaica. Institute for Research on Labor and Employment working paper #142-13.

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.

Hone, L. S. E., and McCullough, M. E. (2020). Are women more likely to wear red and pink at peak fertility? What about on cold days? Conceptual, close, and extended replications with novel clothing colour measures. *British Journal of Social Psychology* **59**, 945–964.

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

Kennedy, L., and Gelman, A. (2021). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *Psychological Methods* **26**, 547–558.

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.

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.

Mischel, W. (1968). *Personality and Assessment*. New York: Wiley.

Rubin, D. B. (1989). A new perspective on meta-analysis. In *The Future of Meta-Analysis*, ed. K. W. Wachter and M. L. Straff, 155–165. New York: Russell Sage Foundation.

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.

Zelner, J., Riou, J., Etzioni, R., and Gelman, A. (2021). Accounting for uncertainty during a pandemic. *Patterns* **2**, 100310.