CHAPTER 17

Visualize as You Randomize

Design-Based Statistical Graphs for Randomized Experiments

Alexander Coppock

Abstract

A good statistical graph for a randomized experiment simultaneously conveys the study's design, analysis, and results. It reveals the experimental design by mapping design elements to aesthetic parameters. It illuminates the analysis by plotting the statistical model in "data-space." When the design and analysis of an experiment are encoded in a plot, the interpretation of the experimental results is clarified. "Analyze as you randomize" is a dictum attributed to Fisher that guides interpretations of experimental data. This chapter extends that principle to visualizations of randomized experiments. While not every experiment requires a visualization, those that do should be visualized in ways that communicate the design and results together.

The purpose of this chapter is to offer a perspective on how to construct statistical graphs for randomized experiments specifically. Many books and articles provide guidance (Chambers et al. 1983; Tufte 1983), inspiration (Lupi and Posavec 2016; W. E. B. Du Bois Center 2018), or technical instruction (Healy 2018; Wickham 2016) for statistical graphs in general. Quite a bit of the data visualization literature is concerned with exploratory data analysis, owing in large part to John Tukey's magnum opus, *Exploratory Data Analysis* (Tukey 1977), which provides procedures for exploring data sets graphically and largely without the aid of computers. The virtues of visualization for exploring data sets for new discoveries have been noted for a long time (e.g., Fisher 1925, p. 25). In experiments, however, the research question to be answered usually flows directly from the design and so does not typically need to be "discovered" through visualization. Indeed, Tukey himself grants that exploratory data analysis is less useful for randomized experiments (Tukey 1977, p. 3). Here, we are not using graphs for exploration; instead, we are using them to reveal experimental design. Much of the common advice about statistical graphs (e.g., "Above all else, show the data" (Tufte 1983) and "Use simplicity in design" (Gordon and Finch 2015)) carries over directly, so in this chapter, I will focus on what is special about graphs for randomized experiments.

The major reason to consider graphs for randomized experiments separately from other visualizations is *experimental design*. Compared with other kinds of studies that seek to measure causal effects, the distinguishing feature of randomized experiments is that the core analytic assumptions can in large part be justified by design, rather than by argument or conjecture. The design-based tradition in the analysis of experiments starts from the design features that are under the researcher's control (who the units are, how the treatment was assigned, how the treatment was delivered, and how the outcomes were measured) and seeks to extract causal inferences while adding a minimum of additional statistical modeling assumptions.

In this chapter, I argue that the visual display of experimental data should follow in the design-based tradition by striving to illuminate experimental design. First, I review the distinctions between design- and model-based descriptive and causal inference. To make explicit what parts of an experimental design a graph can reveal, I rely on the "MIDA" framework for characterizing research designs offered in Blair et al. (2019) and described in brief below. After introducing a minimum of visualization theory, I offer three practical guidelines for graph construction. The chapter concludes with a series of examples that do and do not follow those guidelines.

17.1 Design- and Model-Based Inference, Briefly

The phrase "design-based" has its roots in the survey sampling literature, the main concern of which is drawing descriptive inferences about a population on the basis of a sample. Design-based inference is rooted in the known properties of the sampling procedure: its strata, clusters, and inclusion probabilities. By contrast, model-based statistical inference admits that sometimes the data that have arrived on the analyst's desk did not do so as the result of an explicit sampling procedure – or even if they did, the goal is to make inferences about a population other than the one from which from the sample was drawn. In order to extract inferences about a population on the basis of such a sample, we need to model the process by which the data arose. The description of the two traditions I have given here is necessarily schematic; for lucid discussions of finer distinctions between model- and design-based inference for descriptive quantities, see Little (2004) and Sterba (2009).

The distinction between model-based and design-based inference carries over into causal inference as well. In causal inference, the phrase "design-based" can refer to the idea that the only source of stochasticity in an experiment is the random assignment. Randomization inference is design-based in this sense (Bowers et al. 2013; Gerber and Green 2012; Keele et al. 2012). It is a procedure for conducting hypothesis tests in which the null distribution of a test statistic is explicitly constructed using the set of random assignments to conditions that *could have* occurred according to the experimental design. In Fisher's classic tea-tasting experiment (Fisher 1935, p. 13), a subject (Dr. Muriel Bristol, a biologist working with Fisher at the Rothamsted Experimental Station) made guesses about whether the tea or the milk was added first to eight cups of tea. According to the experimental design, exactly four of the eight cups of tea were assigned to milk first instead of tea first. Fisher simulates the null distribution of the "number correct" test statistic by imagining all $\binom{8}{4} = 70$ ways the milk and tea could have been allocated. He obtains a p-value by comparing the number of cups the subject correctly classified to the null distribution. This p-value does not require any of the extra assumptions required for other hypothesis-testing procedures (e.g., the assumption that, under the null, the test statistic follows an abstract distribution like the t, F, normal, or χ^2 distribution). Beyond keeping modeling assumptions to a minimum, this procedure merits the name "design-based" because it explicitly incorporates the design information that exactly four of eight cups were treated, not two or six.

More loosely, "design-based" can also mean that the core assumptions undergirding experimental inference – random assignment, excludability, and noninterference (Gerber and Green 2012, ch. 2) – are justified by qualitative knowledge of what actually happened in the course of the experiment. We justify the assumption of random assignment on the basis of direct knowledge of assignment mechanism such as the computer script or physical process according to which units were assigned to conditions. We justify excludability by maintaining parallelism in our measurement strategy and designing the treatment to target the relevant theoretical quantity. We justify noninterference by sampling subjects who are far apart from one another in physical and social space. This sense of "design-based" is broader. In this tradition, we root our statistical assumptions in the experimental design and – to the extent possible – refrain from adding extra assumptions.

As an example of model-based analysis of randomized experiments, consider mediation analysis, which is the study of how treatment effects are transmitted through intermediate variables. Consistent estimation of mediation estimands typically requires the additional modeling assumption of sequential ignorability (see Imai et al. 2011 and Chapter 14 in this volume). Under sequential ignorability, the value of the mediator is assumed to be *as-if* randomly assigned within each randomly assigned treatment group, possibly after statistical adjustment. This assumption cannot typically be justified on a design basis and has to be asserted on other grounds that may or may not be convincing to a skeptic (Bullock and Ha 2011). In response to this difficulty, some design-based approaches have been proposed to target mediation estimands while invoking different assumptions (Acharya et al. 2018; Acharya, Blackwell, and Sen 2018; Imai et al. 2013). Other areas in which model-based analysis of experiments is common include addressing attrition and generalizing results to populations outside of the experiment.

17.2 Components of Experimental Designs

To structure the discussion of what components of an experimental design can be visualized, we need to describe what the components of a design are in the first place. In Blair et al. (2019), my colleagues and I offer the "MIDA" framework to characterize research designs. Under that view, an empirical research design consists of four elements: a model, an inquiry, a data strategy, and an answer strategy.

Even when operating in the design-based mode, researchers must rely on a theoretical causal model. These theoretical models inform, but are distinct from, the statistical models described in the previous section. The theoretical model M represents researcher beliefs about the relevant set of exogenous and endogenous variables and their interrelations. For experiments, this usually constitutes researcher beliefs about potential outcomes (how many there are and their distributions).

To fix ideas, consider a two-arm canvassing experiment in which the treatment is a visit from a canvasser making a persuasive appeal to vote for their candidate. In this setting, the theoretical model includes researcher beliefs that each subject only has two potential outcomes (a treated potential outcome and an untreated potential outcome); that the potential outcomes are in a latent probability space; that, on average, the treatment raises subjects' latent probability of supporting the advertising candidate by, for example, 10 percentage points; and that realized outcomes are binary. This model implicitly includes a noninterference assumption that the outcome expressed by a unit depends only on its own treatment assignment and not on the assignments of other units.

The inquiry I is a question about the theoretical model. In experiments, a common inquiry is the average treatment effect

(ATE), defined as the average difference between treated and untreated potential outcomes. There is a tight relationship between inquiries and models. Suppose that, contrary to the noninterference assumption, the outcome expressed by a unit does depend on whether other units are treated, such as whether neighbors talk among themselves about the canvassing visit. In such cases, subjects each have more than two potential outcomes. Even more worrying than bias in our estimates of the ATE is the fact that the very definition of the ATE falls apart because none of the possible potential outcomes can be called "the" treated or untreated potential outcome. If the model changes, then so too must the inquiry. For example, one might redefine the inquiry to be the average difference between being treated versus untreated when all other units are untreated.

The data strategy D is what the researcher does in the world to produce a real data set d. It includes how subjects are brought into the experiment, how they are assigned to treatments, and how outcomes are measured. In our two-arm trial, we might obtain a convenience sample of 500 voters from a proprietary contact list, assign exactly 100 (due to budget or logistical constraints) of them to treatment using complete random assignment, and then use survey data to measure vote intentions. The answer strategy A is the set of statistical procedures we use to generate an answer a using the realized data set. For example, we might use a difference-in-means estimator of the ATE with a t-test for assessing statistical significance. Experimental designs are rarely as simple as two-arm, one-outcome experiments in which the ATE is estimated via difference-in-means, but designs of far greater complexity can usually be straightforwardly characterized in terms of M, I, D, and A as well.

What does it mean for a visualization to reveal the experimental design? The model and the inquiry are imaginary theoretical constructs. The data and answer strategies themselves are procedures that researchers follow. All four features of a design, therefore, are difficult to show in a plot, because imaginary constructs and abstract procedures are hard to represent graphically. Graphs can present the *realization* of the data and answer strategies in ways that illuminate the model and the inquiry. The data strategy D produces data d. The answer strategy A uses data d to produce answer a to inquiry I, which is itself a question about model M. The goal of a design-based statistical graph, therefore, is to visualize d and a in ways that communicate essential features of M, I, D, and A.

17.3 The Semiology and Grammar of Graphics

This section briefly introduces some visualization theory that we need in order to talk about how a graph could successfully communicate experimental design. I will be leaning heavily on a strand of theory that begins with Bertin's "semiology" of graphics (Bertin 1967), runs through Wilkinson's "grammar" of graphics (Wilkinson 2006), and ends with Wickham's development and implementation of the grammar (Wickham 2008, 2016). Bertin's semiology of graphs rests on the idea that the visual signs displayed in a figure encode meaning via a *mapping* from data to aesthetic attributes such as position, size, shape, orientation, brightness, color, and texture. Wilkinson's grammar provides rules according to which the graphs can first be expressed as mathematical objects and can then be projected into a coordinate system. For example, bar charts and pie charts can represent the same mathematical object (the frequency distribution of a categorical variable) in different coordinate systems (Cartesian or polar). The grammar of graphics can be contrasted with a "pen-and-paper" metaphor for creating graphs. With pen and paper, visual elements can be placed on the page in *ad hoc* ways that follow no rules at all, leading to chartjunk and misleading figures (Tufte 1983).

The implication of the semiology and grammar of graphics for the visualization of randomized experiments is that we should map experimental design parameters to aesthetic attributes. For example, we might map the randomized treatment variable to the horizontal axis and the realized outcome variable to the vertical axis. Blocking variables are pretreatment characteristics that describe subsets of subjects among whom separate complete random assignment procedures are carried out. The blocking variable might be mapped to the "facet" aesthetic that controls how the figure will be broken into small multiples. In clustered experiments, whole groups of units are assigned to treatment conditions together. We can map the number of units in each cluster to the aesthetic parameter governing the size of each point. These are all examples of how to express elements of a data strategy in graphical form.

Expressing the answer strategy on a graph is trickier, and to see why, we will adopt the language used by Wickham and coauthors (Wickham 2008; Wickham et al. 2015) to contrast "model-in-data-space" with "data-in-model-space." The word "model" refers here to the statistical model that forms the answer strategy. First and foremost, statistical models applied to data produce data summaries.[1] To summarize the bivariate relationship between two variables X and Y, we often turn to linear regression models that "summarize" Y by providing a guess of the conditional mean of Y at each value of X. It is easy to show this model in data-space. We map the X variable to the horizontal axis and the Y variable to the vertical axis, then we overlay the statistical summary (the linear regression fit) on a scatterplot of Y versus X. This plot is an example of showing the model in data-space.

Showing multiple regression is tougher. Suppose we estimate a regression of the form $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$. We could show a three-dimensional scatterplot with the two-dimensional plane described by β_0, β_1, and β_2 passing through the points, but static three-dimensional plots are difficult to represent on fundamentally flat pages and screens. An alternative is a "res-res" plot (Tukey 1977, ch. 13), in which we estimate two additional regression equations ($Y_i = \alpha_0 + \alpha_1 X_{2,i} + \upsilon_i$ and $X_{1,i} = \gamma_0 + \gamma_1 X_{2,i} + \eta_i$) and plot the residuals from each regression against one another. The best linear summary of the bivariate relationship between the residuals $Y_i^r = Y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 X_{2,i})$ and $X_{1,i}^r = X_{1,i} - (\hat{\gamma}_0 + \hat{\gamma}_1 X_{2,i})$ will be exactly the estimate $\hat{\beta}_1$ obtained from the multiple regression, and a line with this slope can be overlaid, as before, on the residualized data. Notice, however, that taking residuals is a model-based transformation of the data. The res-res plot is an example of showing the data in model-space.

More generally speaking, answer strategies for experiments all have in common that they compare units across randomly formed groups, not within them. For example, the difference-in-means estimator of the ATE is built by constructing estimates of the average treated and untreated potential outcomes out of the realized outcomes expressed in the treatment and control groups, respectively. The ordinary least squares estimator of the ATE does something very similar, but it compares covariate-adjusted estimates of the average treated and untreated potential outcomes.[2] Plotting these statistical models in data-space usually means overlaying the data with estimates of average potential outcomes.

Plotting the actual treatment effect estimates in data-space is difficult, since treatment effects themselves exist only in model-space and not data-space. Because of the Fundamental Problem of Causal Inference (Holland 1986), we cannot directly measure treatment effects, let alone record them in a data set to be visualized. Experimentalists are often in the position of needing to summarize many treatment effect estimates (as in a meta-analysis) or to give descriptive comparisons of treatment effect estimates by subgroups (as in an analysis of treatment effect heterogeneity). Treatment effect estimates can be displayed in a dizzying

[1] See also Gelman (2004), who proposes visualizations of the posterior predictive distribution as a form of model checking. In that setting, the statistical model does more than just summarize the data at hand; it represents a model that could generate new data. Visualizing the posterior predictive distribution is a forceful example of projecting the model into data-space.

[2] For an illuminating discussion of covariate adjustment in randomized experiments, see Lin (2013), especially section 3.

array of formats such as coefficient plots, funnel plots (Light and Pillemer 1984, p. 63), "enhanced" bar plots (Berry and Hauenstein 2017), or scatterplots of original study estimates versus replication study estimates (e.g., Open Science Collaboration 2015, figure 3). These plot formats are not design-based in the sense discussed here. The fact that such plots are not design-based is no criticism; their purpose is often to describe the distribution of a special kind of quantity (causal effects) across many experiments or treatment arms.

17.4 Practical Guidance

The above discussion is somewhat abstract. When trying to apply the design-based approach to my own graphs or when trying to put my finger on what it is that bothers me about others' graphs, I have ambled toward three general guidelines.

First, *invite visual comparisons across randomly formed groups, not across groups formed pre- or post-treatment*. The major distinguishing feature of an experiment is that the causal agent under study was randomly allocated; graphs should emphasize this special design feature. Pretreatment covariates are not randomly assigned, so we should not emphasize comparisons across different values of the covariates. Post-treatment variables are also not randomly assigned. Worse, they are possibly affected by treatment, so comparing units on the basis of post-treatment variables is usually misleading (this is the graphical analogue of post-treatment bias). Graphs typically "invite comparison" by placing units or groups of units closely adjacent to each other in visual space. Randomly formed groups should therefore be kept close to one another and non-randomly formed groups should be separated. Separation can be achieved in a number of different ways, including faceting, differentiating groups with color or shape, or connecting randomly assigned comparisons within nonrandomly assigned groups with lines. Section 17.5.2 offers an illustration of this guideline in the context of a block-randomized experiment.

Second, *show the fitted statistical model (with uncertainty estimates) in data-space*. Following this guideline usually means plotting the data first and laying the statistical model over the data second. Sometimes this means plotting the predictions of the model and not the parameters of the model. Showing uncertainty on a graph can often help communicate important scientific distinctions, such as the distinction between statistical and substantive significance. Showing the model in this way also points out to the viewer that it is our inferences that are uncertain, not the data themselves. When statistical models are visualized alone (i.e., without the underlying data), viewers can be tempted to believe the model is true even when it is not. When the model is shown with the data, models can be shown to be false while nevertheless serving as useful data summaries. Section 17.5.5 demonstrates this guideline when visualizing the interaction of a treatment with a continuous covariate.

Third, *use visual cues like color, shape, diameter, transparency, and facets to reveal design features like blocking, clustering, or differential probabilities of assignment*. Most experiments are not simple two-arm trials with equal probabilities of assignment and a single outcome. The idiosyncratic features of an experiment should be expressed by the visualization. To the extent possible, the graph should highlight *what happened* in the experiment. If groups of units are assigned to treatment as a cluster, the cluster-level outcomes can be represented by points whose diameters are proportional to cluster size (this case is demonstrated in Section 17.5.3). If some units have higher probabilities of assignment than others, the answer strategy will have to account for this design feature. Graphs should account for it as well, possibly by mapping the inverse probability weight to the size or transparency of the point. These visual cues are not just decoration – they point out to the viewer what is special about the specific experiment being visualized.

These guidelines are not hard-and-fast rules. In some settings, they may even be in tension with one another because emphasizing one design feature may come at the cost of obscuring another. Disregarding the first guideline might be appropriate in settings in which the descriptive differences between nonrandomly formed groups are in fact of greater theoretical interest than causal differences across randomly formed groups. See, for example, Kahn (2017), where the baseline differences in behavior between men and women rightly (in my opinion) receive greater attention than the experimentally induced differences in behavior. In my own work on political persuasion, I connect treatment and control subgroup averages with parallel lines to indicate how subgroups of subjects are "persuaded in parallel." Since the (relative) lack of treatment effect heterogeneity is of major theoretical importance in that work, the graphs emphasize the comparison across pretreatment covariates like partisan identification as much as the comparison of treatment to control, in clear violation of the first guideline. The second guideline nudges graph designers to overlay their data summaries on the raw data. But when N is large, plotting each individual data point can be impractical. In such cases, binned summaries (averages of small subsets of the raw data) can serve a similar purpose. Any of three guidelines might be inapplicable in a particular setting. In general, though, how a graph should be constructed will depend on the specifics of the design (i.e., the specifics of M, I, D, and A). Randomized experiments generally share many design features in common, so to the extent a given experiment is similar to the sorts of experiments I describe here, the guidelines will be more applicable.

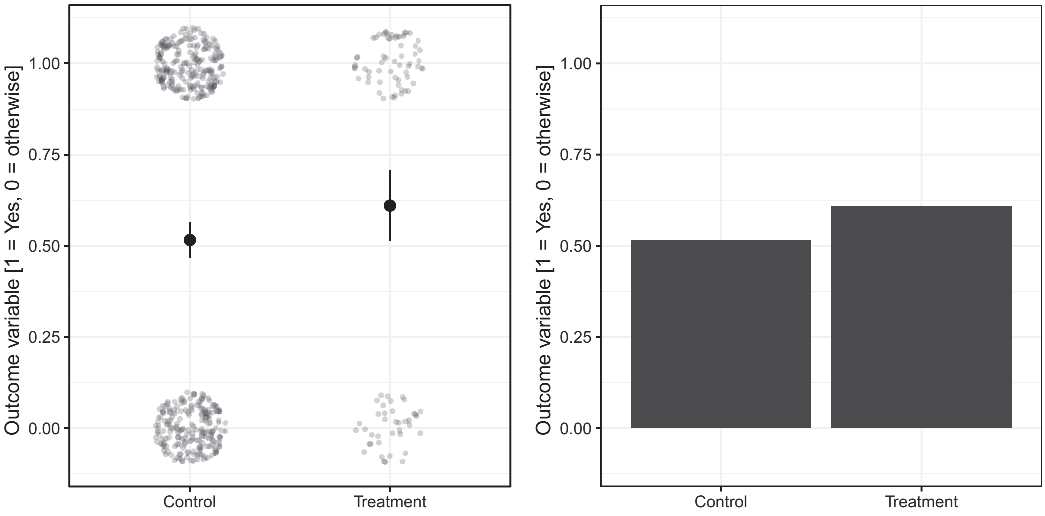
17.5 Example Figures

This section provides examples of strong and weak statistical graphs for randomized experiments. As described in Healy (2018), data visualization advice is often doled out in a "parade of horribles" format, wherein examples of bad graphs are condemned and good graphs are lauded. This more-or-less Manichaean approach emphasizes some graphical features over others. Healy describes three dimensions of visualizations that invite criticism: the aesthetic, the perceptual, and the substantive. I will leave both questions of aesthetics and questions of perception unexplored here. As much as I would like to, I cannot impose my aesthetic tastes on others. I also do not have evidence confirming that my preferred formats actually cause viewers' understanding of experimental designs to improve, though conducting perception experiments in the tradition of Cleveland and McGill (1984) would certainly be feasible. Here, the attention paid to the bad graphs will focus on the substantive design features they misrepresent or fail to communicate. In the examples that follow, all data are fabricated using the R package `DeclareDesign` (Blair et al. 2018) and visualized using the `ggplot2` (Wickham 2016) implementation of the grammar of graphics.[3] In this way, the only "horribles" that are paraded are graphs of my own construction and I can be sure of the design parameters that the graphs are meant to communicate.

17.5.1 Two-Arm Trial

Figure 17.1 takes up the two-arm canvassing experiment example from above. Specifically, imagine a 500-person experiment in which exactly 100 units are treated and the outcome is binary. Figure 17.1a shows data d in a way that recalls data strategy D. The points are distinguished on the horizontal axis according to the realized random assignment. The vertical axis shows the range of the outcome variable and the jittered point clouds remind the viewer that the outcome is binary and offer a sense of the number of units in each condition. The plot also shows pieces of the answer a in a way that recalls answer strategy A. The two group means form the constituent parts of the difference-in-means estimator. The visual comparison of the two group means references the

[3] The programs and data sets used to construct all figures are available on Dataverse at https://doi.org/10.7910/DVN/VE6VSR. Equivalent Stata code for Figure 17.1 is also provided.



(a) Shows the model and data together

(b) Shows the model only

Figure 17.1 ATE simulated two-arm trial. Panel (a) communicates many features of the experimental design, while panel (b) does not.

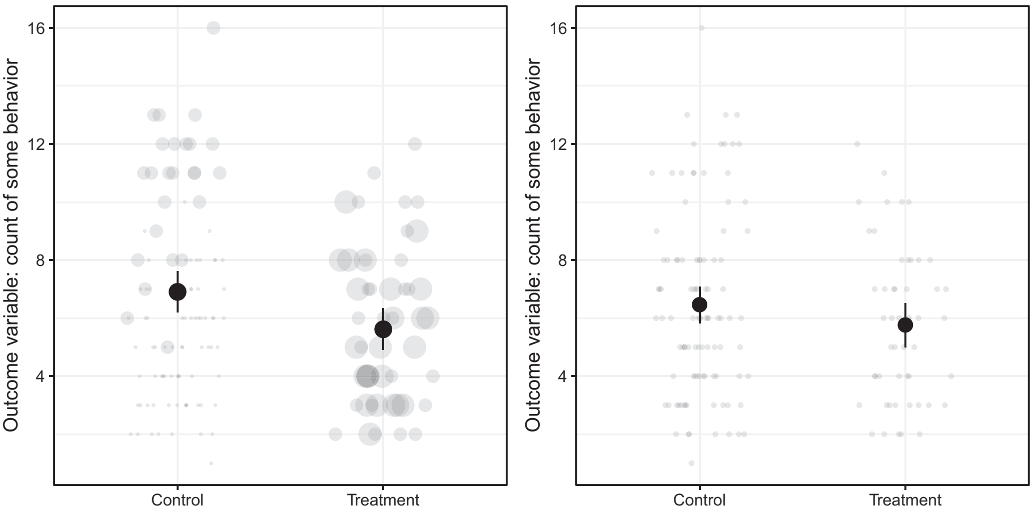
inquiry *I* (the ATE), defined in terms of the model *M* that presumes noninterference and a mapping from latent probabilities to binary measured outcomes. Another virtue of Figure 17.1a is that it displays the uncertainty estimates in the form of 95% confidence intervals for each group mean.

By contrast, Figure 17.1b communicates far less. It only displays two numbers that could be represented just as well in text with no figure at all. Figure 17.1b does not remind the viewer of any important features of the data strategy such as the sample size, the distribution of the outcome variable, or the fraction of units assigned to treatment. It does show some part of the answer strategy (the group means), but it leaves out any measures of uncertainty. In summary, Figure 17.1a follows the second guideline by showing the model and uncertainty in data-space, while Figure 17.1b does not.

17.5.2 Blocked Experiments

In block random assignment, complete random assignment is carried out within separate subgroups of subjects according to their pretreatment covariates. Blocking typically reduces sampling variability relative to complete random assignment and can be seen as a form of covariate adjustment by design. Increasing precision is usually the main reason to block, though logistical constraints may also require some form of blocking. An additional justification for blocking that is sometimes given is the search for treatment effect heterogeneity, on the logic that blocking increases the credibility of heterogeneity analyses. Such analyses by no means require block random assignment and can be carried out whether the assignment procedure is simple, complete, blocked, or otherwise. However, blocking on a variable is credible evidence that the researchers thought the variable was important *ex ante* and that it was not "discovered" through an unprincipled specification search.

For this example, imagine an experiment conducted in two neighborhoods. The first neighborhood has 50 residents and the second has 100. For logistical reasons, the experimental partner needs to treat exactly 25 residents in each neighborhood. This data strategy induces differential probabilities of assignment. The probability of treatment is higher in the first neighborhood than the second, which could cause bias in the unadjusted difference-in-means estimator. The answer strategy must account for the probabilities of assignment somehow.



(a) Point size is proportional to the inverse probability weights and weighted means are plotted

(b) Differential probabilities of assignment are ignored, yielding a biased comparison of group averages

Figure 17.2 A simulated block-randomized experiment. Panel (a) incorporates the differential probabilities of assignment induced by the blocking, while panel (b) does not.

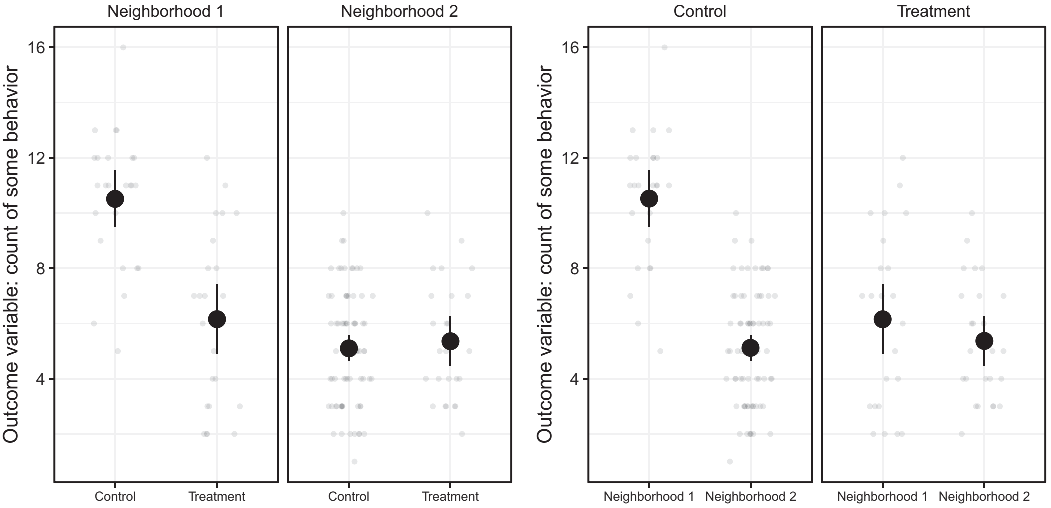
Approaches include conditioning on blocks via a stratified estimator, controlling for blocks in a regression setting, or weighting units by the inverse of the probability of being in the treatment condition to which they were assigned. The outcome is a count of some behavioral outcome over the course of the experiment, which varies by neighborhood and by treatment condition.

Figure 17.2 shows two visualizations of the blocked experiment. In Figure 17.2a, the diameters of the points representing residents are proportional to the inverse probability weights. Additionally, the weighted group means are overlaid. The visualization in Figure 17.2b ignores the differential probabilities of assignment altogether: the points are all the same size and the unweighted group means are plotted. In this cooked-up example, the correct analysis shows that the ATE estimate is large and negative whereas the incorrect analysis estimates the ATE to be close to zero. Incorporating design information into the estimation of treatment effects is obviously important, and Figure 17.2 shows that it is equally important to include such information in visualizations.

Next, we consider treatment effect heterogeneity by block. Both panels of Figure 17.3 follow the design guidance given in Figure 17.1. The panels differ, however, by the variable that creates the facets. In Figure 17.3a, we facet by block. Within each facet, the groups that are compared are formed by random assignment: we see small effects of treatment in neighborhood 1 and large negative effects of treatment in neighborhood 2. By contrast, in Figure 17.3b, we facet by randomly assigned group, so we compare across schools and within treatment group. Figure 17.3a follows the first piece of graphical design advice – *invite visual comparisons across randomly formed groups, not across groups formed pre- or post-treatment* – while Figure 17.3b does not. The reasons to prefer Figure 17.3a in blocked settings carry over immediately to any search for treatment effect heterogeneity by pretreatment covariates, regardless of whether the covariates were used to create blocks.

17.5.3 Clustered Experiments

In cluster-randomized trials, whole groups of subjects are assigned to treatment conditions together. Clustering causes some inferential problems – it decreases precision and can even induce bias in standard analytic



(a) Emphasizes within-block comparison (b) Emphasizes across-block comparison

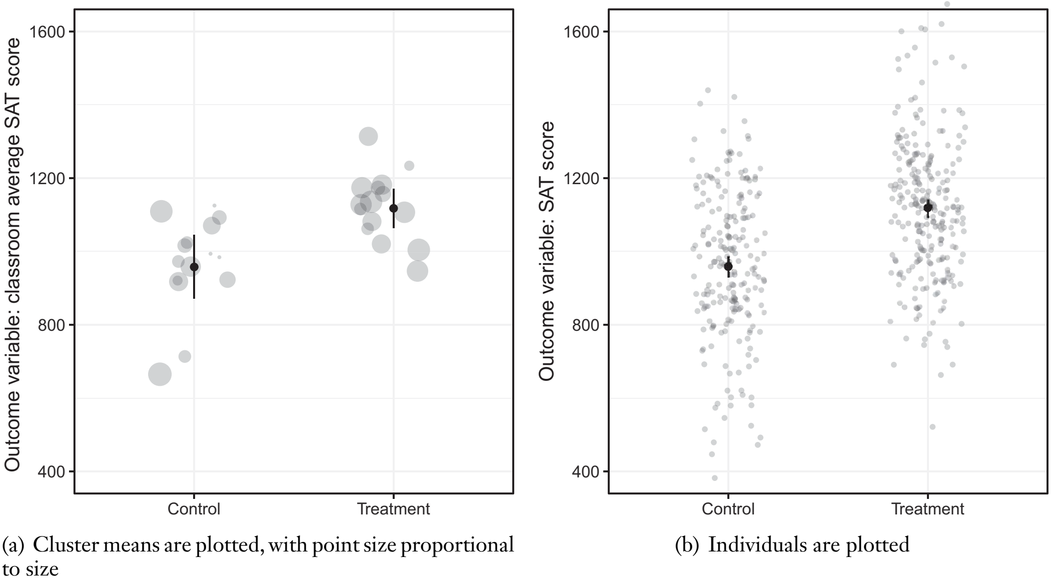
Figure 17.3 The same simulated block-randomized experiment. Panel (a) compares across the randomly assigned partitions, while panel (b) does not.

procedures (Imai et al. 2009; Middleton 2008). Clustering should usually be avoided where possible, but sometimes logistical, budget, or theoretical constraints require treatments to be allocated at the cluster level. A prototypical example is the classroom-level experiment. Because whole classrooms of students receive instruction together, education interventions are often implemented at the classroom level and outcomes are measured at the student level. In this example, we imagine a "test-prep" treatment that is randomly assigned to some classrooms but not to others; the outcome is measured on a 400–1600-point scale. The analysis must account for the clustering, either by using clustered standard errors or by aggregating outcomes to the cluster-level and weighting the difference-in-means estimator by cluster size.

Two ways of plotting the data and summaries from this experiment are shown in Figure 17.4. Figure 17.4a plots cluster means with diameters proportional to the cluster size. The weighted means are plotted along with 95% confidence intervals that are calculated accounting for clustering. The main virtue of this visualization is that it insists on the clustered randomization procedure. By contrast, Figure 17.4b does not communicate the randomization procedure because it falsely gives the impression that all of the individual points are independent. While the group means are the same in both panels, the 95% confidence intervals in Figure 17.4b were constructed ignoring the clustering. As a rule, if the answer strategy requires clustering in order to obtain good variance estimates, the visualization should somehow communicate clustering as well.

17.5.4 Covariate Adjustment

Experimentalists often include pretreatment covariates in their statistical models to increase the precision of their estimates. Gerber and Green (2012, ch. 4) draw a connection between covariate adjustment and differencing off a pretest score that provides a nice intuition for why covariate adjustment helps to increase precision. Differencing off the pretest score removes a large portion of the idiosyncratic variation in the post-test score; it is also equivalent to *setting* the regression coefficient on the covariate to 1. If adjustment is carried out using regression, the coefficient is instead chosen by the model fitting algorithm, which tends to offer larger precision gains than differencing.



(a) Cluster means are plotted, with point size proportional to size

(b) Individuals are plotted

Figure 17.4 A simulated cluster-randomized experiment. Panel (a) emphasizes the cluster-randomized design and displays confidence intervals that account for clustering, while panel (b) does neither.

Similarly, visualization can help convey to viewers how the answer strategy (regression) sharpens estimates. One way to do this is to present the unadjusted and adjusted data side by side, as shown in Figure 17.5. In the "unadjusted" facet, the data are plotted with the unadjusted regression line plotted on top. The adjusted plot is a res-res plot (discussed above) in which both the outcome variable and the treatment variable have been residualized by the pretreatment covariates. Here, I use the procedure recommended in Lin (2013) to adjust each treatment arm separately, equivalent to interacting the covariates with the treatment indicator and then predicting the average outcome in each treatment condition. In order to clearly see the variance reduction from covariate adjustment, the vertical scale of both facets (but not their range) is set to be the same. Typically, this scale must be set manually, since most visualization software will alter the scale to fit the range of the data. A common piece of advice for experimentalists is to present treatment effect estimates with and without covariate adjustment side by side in a regression table so that readers can appreciate the work that covariate adjustment is doing for the analysis. Figure 17.5 is the graphical analogue of that side-by-side comparison.

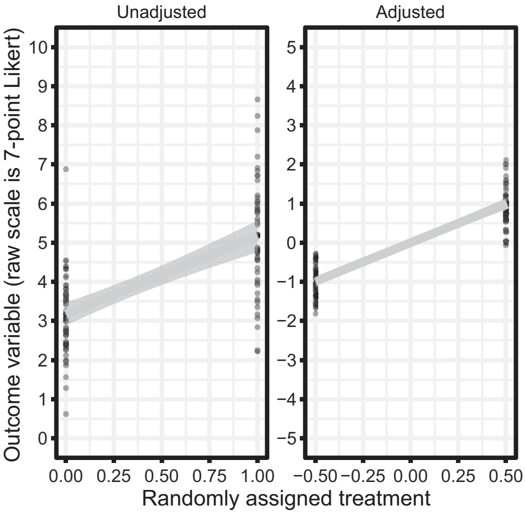
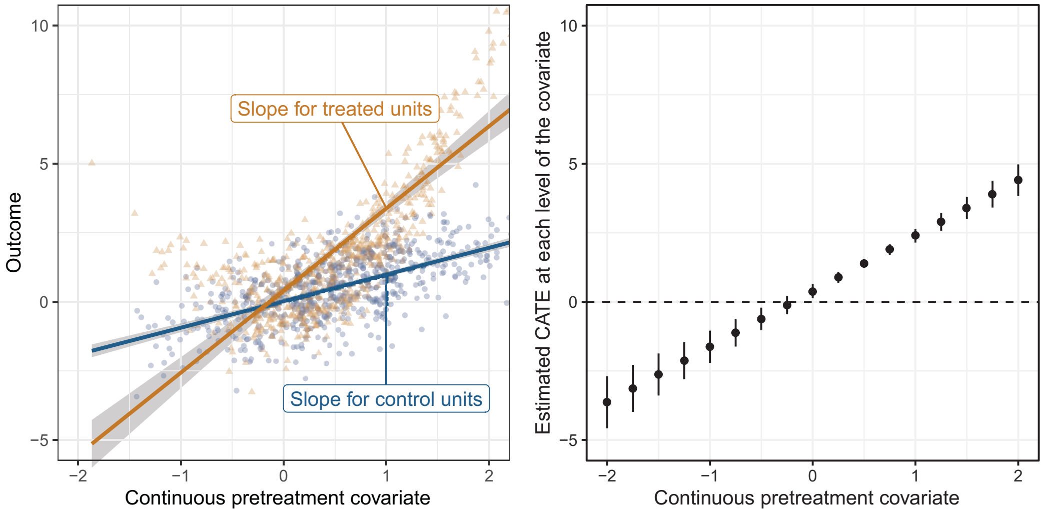


Figure 17.5 Covariate adjustment. The contrast between the two facets shows how the adjustment increases statistical precision.

17.5.5 Interactions with a Continuous Covariate

Another way to incorporate covariates into experimental analysis is to estimate



(a) Shows the model and data together

(b) Model predictions masquerade as data

Figure 17.6 Simulated experiment with a continuous pretreatment covariate. Both panels represent the same statistical model, but panel (a) plots the model and data together in data-space, while panel (b) plots the model only in model-space.

conditional average treatment effects (CATEs). Figure 17.3a shows one way to visualize CATEs when the "condition" is a categorical pretreatment covariate. When covariates are continuous, however, the one-category-per-facet small multiples approach will not work. Continuous covariates require a statistical model that predicts the CATE at each level of the covariate. These models can be flexible or rigid, theoretically motivated or data-driven, depending on the research setting. The most common model is by far an ordinary least squares regression with a treatment-by-covariate interaction. This model is equivalent to fitting separate (straight) lines to the treated and untreated units. The CATE estimate at each level of the covariate is the difference between the fitted lines.

Figure 17.6a shows a design-based way to visualize a continuous interaction. The outcome is mapped to the vertical axis and the pretreatment covariate is mapped to the horizontal axis. The randomly assigned treatment is mapped to the shape aesthetic, with triangles for treated units and circles for untreated units. We overlay the fitted statistical model so that we can see the model in data-space. The graph shows that, according to the statistical model, the treatment effect is different at different levels of the covariate. The model estimates negative effects for low values of the covariate and positive effects for high values of the covariate. The graph also reveals that the statistical model does not fit the data particularly well, especially at low values of the covariate. The graph indicates that the answer strategy of fitting a linear treatment-by-covariate interaction will yield misleading CATE estimates for low values of the covariate.

Figure 17.6b visualizes the results by plotting the estimated CATE for a series of values of the covariate. Graphs like Figure 17.6b are often presented because they explain complex statistical models much more clearly than the corresponding regression tables (Brambor et al. 2006) and because some popular software packages make them easy to produce (e.g., King et al. 2000). However, such graphs can convey a false sense of certainty. The confidence intervals are computed conditional on the statistical model being correct, but the model is not correct (for diagnostic checks that can alert analysts to mispecifications in interactive models, see Hainmueller et al. 2019). The separate points give the impression that each point is independent, but that is not true. The model predictions masquerade as data; they mislead the viewer about the relative contribution of the data and the statistical model to the overall inferences. I want to emphasize that the problem with Figure 17.6b is not just that the statistical model is incorrect, but also that it obscures the experimental design. The statistical model shown in Figure 17.6a is equally wrong – both panels represent the same model, which could of course be improved by more flexible regression specification. The main virtue of Figure 17.6a over Figure 17.6b is that, by displaying the model in data-space, we understand the work it does to summarize the data.

17.5.6 Noncompliance

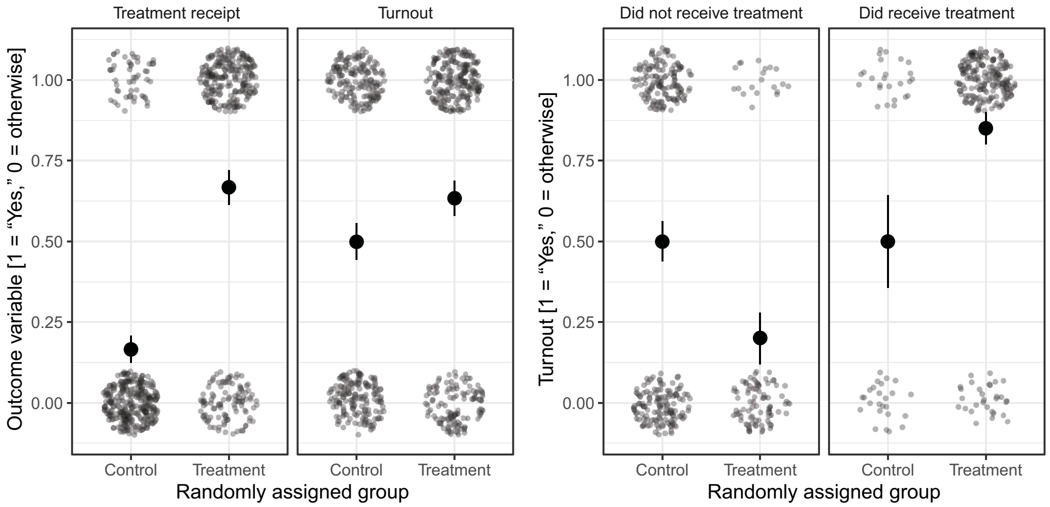
Experiments encounter noncompliance when subjects assigned to the treatment group do not receive treatment or when control group subjects take treatment. The standard analytic approach in such cases is to estimate the effect of assignment on two post-treatment variables: treatment receipt and the outcome of interest. Under some additional assumptions, the ratio of these two intention-to-treat estimates forms an estimate of the complier average causal effect, sometimes called the local average treatment effect (LATE; Angrist et al. 1996). Many alternative analyses (e.g., a "per-protocol" or an "as-treated" analysis) condition in some way on treatment *receipt*. Because receipt is a post-treatment variable, such analyses are prone to post-treatment bias and do not typically produce good estimates of any policy or theory-relevant estimands. For a textbook introduction to the analysis of experiments under noncompliance, see Gerber and Green (2012, chs. 5 and 6).

Experimentalists must take care to ensure their visualizations also avoid conditioning on post-treatment variables. Figure 17.7a successfully avoids this error. It follows the format of Figure 17.1 to show the group means for two different outcomes according to the random assignment. Figure 17.7b conditions on a post-treatment variable by displaying units that did and did not receive treatment in separate facets. The left facet compares units that did not end up taking treatment on the basis of the assignment; the right facet does the same for units that did take treatment. The resulting group means in Figure 17.7b are mostly useless for causal inference, as neither the within-facet nor the across-facet comparisons are informative. Figure 17.7a follows the first guideline (invite comparisons across randomly formed groups), while Figure 17.7b does not. This principle extends to other post-treatment variables beyond treatment receipt, such as manipulation checks (Aronow et al. 2019).

17.5.7 Attrition

Experiments sometimes encounter attrition, or missingness in the outcome variable. Conditioning the analysis on non-missingness (i.e., dropping units for which the outcome is missing) can induce bias because non-missingness is a post-treatment variable. One alternative that does not condition on any post-treatment variables is extreme value bounds (Manski 1999). This approach sidesteps the problem by computing best-case and worst-case bounds. An upper bound on the treatment effect estimate is obtained by imputing the maximum possible outcome for all missing treated units and the minimum possible outcome for all missing control units. The lower bound is obtained by doing the reverse. These bounds are the logical limits on the possible values for the ATE estimate that are consistent with the observed data and knowledge of the minimum and maximum possible values of the outcome. These bounds themselves are estimates and are subject to uncertainty due to sampling variability.

As an answer strategy, extreme value bounds can be difficult to communicate to readers because, even though they rely on a relatively simply logic, they are nevertheless unfamiliar to many. Figure 17.8 shows one way to visualize the procedure in a design-based graph. As in Figure 17.1, the random assignment is mapped to the horizontal axis and the outcome variable (measured on a seven-point Likert scale) is mapped to the vertical axis. Whether the data are observed or imputed is mapped to both color and



(a) Does not condition on post-treatment variables

(b) Conditions on a post-treatment variable (treatment receipt)

Figure 17.7 Simulated experiment encountering two-sided noncompliance. Panel (a) compares randomly formed groups on two different dependent variables, while panel (b) conditions on a post-treatment variable by displaying those who did and did not receive treatment in separate facets.

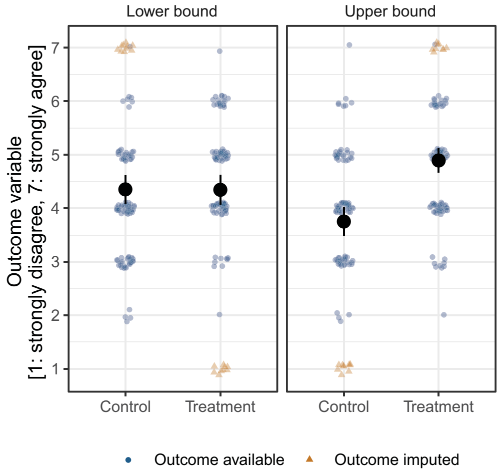


Figure 17.8 A simulated experiment encountering attrition. The panels show observed and imputed data under the worst-case and best-case scenarios.

shape: the observed data are shown in blue circles and the imputed data are shown as gold triangles. This double aesthetic mapping ensures legibility regardless of whether the graph is viewed in color or in grayscale.

The lower bound facet shows the worst-case scenario. The imputed values for the control group are as high as possible, whereas the imputed values for the treatment group are as low as possible. In this case, the resulting group means (shown as large black points) are very similar in treatment and control, so the implied worst-case treatment effect estimate is very close to zero. In the upper-bound facet, however, the imputations are reversed. Now the imputations in the control group are as low as possible and the imputations in the treatment group are as high as possible. The upper-bound treatment effect estimate is close to a full-scale point. These extreme value bounds characterize how big or small the treatment effect could be under the worst- and best-case scenarios: even with attrition that is extremely correlated with potential outcomes, the ATE is likely to lie somewhere between zero and one. The visualization helps to communicate this somewhat convoluted line of argumentation by distinguishing what is observed and what is imputed.

17.6 Discussion

Failure to incorporate design information into analyses of experimental data can lead to inferential errors of every variety.

Failure to account for clustering can lead to overconfidence; failure to account for differential probabilities of assignment can lead to bias; and failure to differentiate clearly between which variables were and were not randomly assigned can lead to misinterpretation on the part of readers, if not the analysts themselves.

In this chapter, I have argued that experimenters should apply similar care to their visualizations. Clustering can be accounted for by plotting points with diameters proportional to cluster size; differential probabilities of assignment induced by blocking can be addressed though faceting by block; and visualizations can emphasize divisions based on treatment assignment over divisions based on pre- or post-treatment variables. Following the principle of "visualize as you randomize" means holding visualizations of experimental data to the same standards to which we hold analyses of experimental data.

Some of the advice in this chapter could apply to visualizations of nonexperimental studies as well. Observational research designs can also be expressed in terms of M, I, D, and A (see Blair et al. 2019, pp. 846–849 for examples), and good visual summaries of such designs will communicate these essential features. For example, a study using an instrumental variables design might use a display similar to Figure 17.7a to show both the first-stage and the reduced-form analyses. Visualizations of regression discontinuity designs already often follow the second piece of design advice to plot the data (or small aggregations of the data in "bins") and model together (see, e.g., Klašnja and Titiunik 2017, figure 1). Time-series graphs have often been deployed to bolster the parallel trends assumption in difference-in-difference designs; many such graphs could be improved by showing the fitted models in data-space (e.g., Pischke 2007, figure 2). The identification strategies of many observational studies turn on a claim that the observed data (possibly after adjustment) are similar to the sort of data that would arise from a randomized experiment. Design-based statistical graphs for observational studies can help readers understand how exactly the analogy to experiments is being drawn and whether the analogy is strong enough to license causal inferences.

The foregoing examples of good and bad visualizations are not meant as templates, nor should anyone feel compelled to make the same aesthetic choices as I have above. Different designs will call for different visualizations. The main purpose of the examples is to be concrete about the otherwise amorphous idea of mapping design parameters to aesthetic parameters and to show how graphs might succeed or fail at doing so.

All of the examples were conducted in R, which is a popular programming language in 2019 (the time of writing). I am sure that, in the near future, many new ways to create statistical graphs will be invented, so the precise procedures for making these particular graphs will become obsolete. When choosing a language in which to create graphs, analysts should be sure to be able to create facets for small multiples and to be able to overlay summaries on top of raw data. These tasks are far more difficult in some languages than others, but we should not allow our programming languages to restrict what sort of figures we create.

Finally, I provide a brief reflection on the many purposes of graphs for randomized experiments. First and foremost, they are included in academic journal articles and books in order to elucidate experimental design and bolster scientific conclusions. But academic outlets are not the only place for design-based experimental graphs. In my experience, they far outpace tabular[4] presentations of experimental results in academic presentations, especially in settings when the audience does not have access to the crucial design information that a graph could communicate. Graphs serve a similar purpose for communicating with nonacademic audiences. In my view, the

[4] Not a little ink has been spilled on the topic of tables versus graphs (Gelman 2011; Gelman et al. 2002; Kastellec and Leoni 2007). In the age of online appendices, I think the argument can mostly be put to the side: the statistical summaries that we show in graphs can easily be reproduced in tabular format as well.

purpose of a visualization is not just to explain *what* we think we know, but also *why* we think we know it. Since the answer to the second question is "experimental design," the argument for design-based graphs becomes self-evident.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26(4): 357–378.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444–455.

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects who Fail a Manipulation Check." *Political Analysis* 27(4): 572–589.

Berry, William D., and Matthew Hauenstein. 2017. "Merging Graphics and Text to Better Convey Experimental Results: Designing an 'Enhanced Bar Graph'." *PS: Political Science & Politics* 50(3): 831–836.

Bertin, Jacques. 1967. *Sémiologie Graphique*. Paris: Editions Gauthier-Villar. English translation by W. J. Berg as *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI: University of Wisconsin Press, 1983. Reissued in 2010, Ersi Press.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113(3): 838–859.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Neal Fultz. 2018. "DeclareDesign." Software package for R. URL: http://declaredesign.org

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. "Reasoning about Interference Between Units: A General Framework." *Political Analysis* 21(1): 97–124.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1): 63–82.

Bullock, John G., and Shang E. Ha. 2011. "Mediation Analysis Is Harder than It Looks." In *Cambridge Handbook of Experimental Political Science*, Cambridge, UK: Cambridge University Press, p. 959.

Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey. 1983. *Graphical Methods for Data Analysis*. Boston, MA: Wadsworth.

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.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Gelman, Andrew. 2004. "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13(4): 755–779.

Gelman, Andrew. 2011. "Why Tables Are Really Much Better than Graphs." *Journal of Computational and Graphical Statistics* 20(1): 3–7.

Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *American Statistician* 56(2): 121–130.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Gordon, Ian, and Sue Finch. 2015. "Statistician Heal Thyself: Have We Lost the Plot?" *Journal of Computational and Graphical Statistics* 24(4): 1210–1229.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2): 163–192.

Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–960.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1): 5–51.

Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24(1): 29–53.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black

Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4): 765–789.

Kahn, Sarah. 2017. "Personal is Political: Prospects for Women's Substantive Representation in Pakistan." Unpublished manuscript.

Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. "Using Graphs instead of Tables in Political Science." *Perspectives on Politics* 5(4): 755–771.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56(2): 484–499.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 347–361.

Klašnja, Marko, and Rocio Titiunik. 2017. "The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability." *American Political Science Review* 111(1): 129–148.

Light, Richard J., and David B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics* 7(1): 295–318.

Little, Roderick J. 2004. "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99(466): 546–556.

Lupi, Giorgia, and Stefanie Posavec. 2016. *Dear Data*. San Francisco, CA: Chronicle Books.

Manski, Charles F. 1999. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Middleton, Joel A. 2008. "Bias of the Regression Estimator for Experiments Using Clustered Random Assignment." *Statistics & Probability Letters* 78(16): 2654–2659.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): aac4716.

Pischke, Jörn-Steffen. 2007. "The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years." *Economic Journal* 117(523): 1216–1242.

Sterba, Sonya K. 2009. "Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research* 44(6): 711–740.

Tufte, Edward. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

W. E. B. Du Bois Center. 2018. *WEB Du Bois's Data Portraits: Visualizing Black America*. San Francisco, CA: Chronicle Books.

Wickham, Hadley Alexander. 2008. "Practical tools for exploring data and models." PhD thesis, Iowa State University.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer.

Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. "Visualizing Statistical Models: Removing the Blindfold." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8(4): 203–225.

Wilkinson, Leland. 2006. *The Grammar of Graphics*. Berlin: Springer Science+Business Media.