The Review *of* Economics *and* Statistics

VOL. CVII MAY 2025 NUMBER 3

FACTORIAL DESIGNS, MODEL SELECTION, AND (INCORRECT) INFERENCE IN RANDOMIZED EXPERIMENTS

Karthik Muralidharan, Mauricio Romero, and Kaspar Wüthrich*

Abstract—Factorial designs are widely used to study multiple treatments in one experiment. Although t-tests using a fully saturated "long" model provide valid inferences, "short" model t-tests (that ignore interactions) yield higher power if interactions are zero, but incorrect inferences otherwise. Of 27 factorial experiments published in top-five journals (2007–2017), nineteen use the short model. After including interactions, more than half of their results lose significance. Based on recent econometric advances, we show that power improvements over the long model are possible. We provide practical guidance for the design of new experiments and the analysis of completed experiments.

I. Introduction

CROSS-CUTTING or factorial designs are widely used in field experiments. For example, 27 out of 124 field experiments published in top-five economics journals during 2007–2017 use cross-cutting designs. One rationale is that the power for detecting main treatment effects is higher if interactions between treatments are ignored in estimation and inference (with the implicit assumption that interactions are zero or negligible). This can make factorial designs a cost-effective way of studying multiple treatments.[1] A second rationale is to "explore" if there are meaningful interactions across treatments. This paper is motivated by the observation that both of these rationales can be problematic in practice.

To fix ideas, consider a setup with two randomly assigned binary treatments. The researcher can estimate either a fully saturated "long" model (with dummies for both treatments and their interaction) or a "short" model (only including dummies for both treatments). The long model yields consistent estimators for the main treatment effects of both treatments and is always correct for inference regardless of the true value of the interaction effect. However, if the true value of the interaction effect is zero, the short model yields consistent estimators and has greater power for conducting inference on the main effects.

The power gains from the short model, however, come at the cost of an increased likelihood of incorrect inference relative to a business-as-usual counterfactual (defined as outcomes in a pure experimental control group) if the interaction effect is not zero. Out of 27 field experiments published in top-five economics journals during 2007–2017 using cross-cutting designs, nineteen (over 70%) do not include all interaction terms in the main specifications. We reanalyzed the data from these papers by also including the interaction terms.[2] Doing so has nontrivial implications for inference on the main treatment effects. The median absolute value of the change in the point estimates is 96%, about 26% of estimates change sign, and 53% (29 out of 55) of estimates reported to be significant at the 5% level are no longer so after including interactions. Even if we reanalyze only "policy" experiments, 32% of the estimates (six out of nineteen) are no longer significant after including interactions.[3]

In practice, researchers often estimate the long model first and test if the interaction is significant, and then focus on the short model if they do not reject that the interaction is zero. However, such data-dependent model selection leads to invalid inferences (Leeb & Pötscher, 2005, 2006, 2008; Kahan, 2013) and should thus be avoided. Further,

Received for publication January 10, 2022. Revision accepted for publication December 28, 2022. Editor: Xiaoxia Shi.

*Muralidharan: UC San Diego; NBER; J-PAL; Romero: ITAM; J-PAL; Wüthrich: University of Michigan; CESifo.

We are grateful to Isaiah Andrews, Tim Armstrong, Prashant Bharadwaj, Arun Chandrasekhar, Clement de Chaisemartin, Gordon Dahl, Stefano DellaVigna, Esther Duflo, Graham Elliott, Andrew Gelman, Markus Goldstein, Macartan Humphreys, Guido Imbens, Hiroaki Kaido, Lawrence Katz, Michal Kolesar, Adam McCloskey, Craig McIntosh, Rachael Meager, Paul Niehaus, Ben Olken, Gautam Rao, Andres Santos, Jesse Shapiro, Diego Vera-Cossio, and many seminar participants for comments and suggestions. We are also grateful to the authors of the papers we reanalyze for answering our questions and fact checking that their papers are characterized correctly. Finally, we would like to thank Tim Armstrong, Adam McCloskey, Graham Elliott, Michal Kolesar, and Soonwoo Kwon, who graciously answered questions about the econometric methods they developed and how to implement them. Sameem Siddiqui provided excellent research assistance. All errors are our own. Financial support from the Asociación Mexicana de Cultura, A.C. is gratefully acknowledged by Romero.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01317.

[1] As Kremer (2003) puts it: "Conducting a series of evaluations in the same area allows substantial cost savings…. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically…. This tactic can be problematic, however, if there are significant interactions between programs."

[2] The full list of 27 papers is in table A1. We reanalyzed fifteen out of the nineteen that do not include all interactions in the main specification. The other four papers did not have publicly accessible data.

[3] We define a policy experiment as one which studies a program or intervention that could be scaled up, as opposed to a conceptual experiment, which aims to test for the existence of facts or concepts such as discrimination (e.g., résumé audit experiments).

The Review of Economics and Statistics, May 2025, 107(3): 589–604

https://doi.org/10.1162/rest_a_01317

cross-cutting experiments are rarely adequately powered to detect meaningful interactions (see section IIF). Thus, this two-step procedure will almost always fail to reject that the interaction term is zero, even when it is different from zero. As a result, the rate of incorrect inference using this two-step model-selection procedure will continue to be nearly as high as that from just running the short model.

The lack of power to detect interactions combined with a focus on statistical significance also makes it challenging to use factorial designs to "explore" whether interactions are meaningful. The interaction estimator's variance is always larger than that of the main effects estimators, making the sample size requirements for detecting interactions much more onerous.[4] This leads to most factorial experiments being underpowered to detect interactions. As a result, point estimates of interactions will on average substantially overstate the true effect, *conditional on being significant*. This problem has been referred to by Gelman and Carlin (2014) as Type-M error.

Textbook treatments of factorial designs (Cochran & Cox, 1957; Gerber & Green, 2012) and guides to practice (Kremer, 2003; Duflo et al., 2007) are careful to clarify that treatment effects using the short model should be interpreted as either (a) being conditional on the distribution of the other treatment arms in the experiment or (b) as a composite treatment effect that includes a weighted-average of the interactions with other treatments. However, as we argue in section IIC, this weighted average is a somewhat arbitrary construct, can be difficult to interpret in high-dimensional factorial designs, and is typically neither of primary academic interest nor policy-relevant. Consistent with this view, none of the nineteen experimental papers that focus on the short model motivate their experiment as being about estimating a weighted-average treatment effect.

The status quo of focusing on the short model is problematic for at least three reasons. First, ignoring interactions affects internal validity against a "business-as-usual" counterfactual. If the interventions studied are new, the other programs may not even exist in the study population. Even if they do, there is no reason to believe that the distributions in the population mirror those in the experiment. Thus, to the extent that estimation and inference of treatment effects depend on what *other* interventions are being studied in the same experiment, ignoring interactions is a threat to internal validity.

Second, "absence of evidence" of significant interactions may be erroneously interpreted as "evidence of absence." The view that interactions are second order (as implied when papers only present the short model) may have been influenced partly by the lack of evidence of significant interactions in most experiments to date. However, as we show in section IIF, this is at least partly because few experiments are adequately powered to detect meaningful interactions. There is now both experimental (Duflo et al., 2015; Mbiti et al., 2019) and nonexperimental (Kerwin & Thornton, 2021; Gilligan et al., 2022) evidence that interactions matter. Indeed, a long tradition in development economics has highlighted the importance of complementarities across programs in alleviating poverty traps (Ray, 1998; Banerjee & Duflo, 2005), which suggests that assuming away interactions in empirical work may be a mistake.

Third, there is well-documented publication bias toward significant findings (e.g., Franco et al., 2014; Andrews & Kasy, 2018; Christensen & Miguel, 2018; Abadie, 2020). This can also affect evidence aggregation because meta-analyses and evidence reviews often include only published studies. Thus, the sensitivity of the significance of main effect estimates to the inclusion/exclusion of interaction terms (which we document in this paper) is likely to have nontrivial implications for how evidence is published, summarized, and translated into policy.

Having documented the limitations of the short model, we consider if it is possible to improve power relative to the long model *while maintaining size control* for relevant values of the interactions. The two-sided long model t-test is the uniformly most powerful unbiased test (e.g., van der Vaart, 1998; Elliott et al., 2015a). This result implies that if one insists on size control for *all* values of the interaction effect, any procedure that is more powerful than the t-test for some values of the interactions must have lower power somewhere else. This classical result motivates imposing restrictions on the interaction effects based on prior knowledge to improve power. We explore three different approaches.[5]

The first approach, based on Elliott et al. (2015a), is a nearly optimal test that targets power toward an a priori likely value of the interaction (e.g., a value of zero), while controlling size for *all* values of the interaction. This approach comes close to achieving the maximal theoretically possible power near the likely value of the interaction but exhibits lower power than the long model t-test farther away. We then consider two approaches based on Armstrong et al. (2020) and Imbens and Manski (2004) for constructing confidence intervals for the main effects under restrictions on the magnitude of the interactions based on prior knowledge. When the prior knowledge is correct, these approaches control size and yield substantial power gains relative to the long model t-tests. However, these power gains come at the cost of size distortions if the prior knowledge is incorrect.

Based on the analysis above, we recommend—in the interest of transparency—that factorial experiments report results from the long regression model (even if only in an appendix). Long model t-tests are easy to compute even in complicated factorial designs and have appealing optimality properties. Further, the justification for omitting

[4] For example, one would need an eight times larger sample to detect an interaction than to detect a main effect when the interaction is half the size of the main effect; see section IIF and appendix A.3.

[5] In appendix A.6, we explore a fourth approach based on McCloskey (2017, 2020), which is based on a Bonferroni-type correction after consistent model selection.

interactions should *not* be that these were not significant in the long model (because of the model selection issue discussed above). Rather, if researchers would like to focus on results from the short model, they should clearly indicate that treatment effects should be interpreted as composite effects that include a weighted average of interactions with other treatments (and specify the estimand of interest in a preanalysis plan). This will enable readers to assess the extent to which other treatments may be typical background factors that can be ignored.

For the design of new experiments, if the primary parameters of interest are the main effects, a natural alternative is to leave the "interaction cells" empty and increase the number of units assigned to the main treatment(s) or the control group. Our simulations show that this design-based approach yields more power gains than the econometric methods discussed above for most of the relevant values of the interaction.

Reviewing classic texts on experimental design, we identify four cases where factorial designs and analyses of the short model may be appropriate. The first is where the goal is to explore several treatments efficiently to identify promising interventions for *further* testing (e.g., Cochran & Cox, 1957). However, most policy experiments are run only once, making factorial designs and short model estimates less desirable.

The second is when the goal is not to test whether a given treatment has a "significant" effect, but to minimize mean squared error (MSE) criteria (or other loss functions) involving a bias-variance trade-off in estimating the main effects (e.g., Blair et al., 2019). However, a key rationale for experimental evaluations of policies and programs is to generate unbiased estimates, making the bias in the short model unattractive.

The third is to improve external validity. Cochran and Cox (1957, p. 152) recommend bringing in subsidiary factors into factorial designs to test main effects over a wide range of conditions; also see Fisher (1992). Thus, factorial designs and analyses of the short model may be fine when one dimension of the experiment is studying reasonable variants of the main treatment, but less so when all treatments are of primary interest.

The fourth is the case of conceptual (as opposed to policy) experiments, such as résumé audit studies, where many of the characteristics that are randomized (such as age, education, race, and gender) do exist in the population. When feasible, we recommend having the treatment share of various characteristics being studied be the same as their population proportion. Doing so will make the short-model coefficient more likely to approximate a population relevant parameter of interest. We discuss each of these four rationales along with relevant examples in section V.

Our first contribution is to the literature on the design of field experiments. Bruhn and McKenzie (2009), List et al. (2011), and Athey and Imbens (2017) provide guidance on the design of field experiments, but do not discuss when and when not to implement factorial designs. Duflo et al. (2007, p. 3932) implicitly endorse the use of factorial designs by noting that they "[have] proved very important in allowing for the recent wave of randomized evaluations in development economics."

Our reanalysis of existing experiments as well as simulations suggest that *there is no free lunch*. The perceived gains in power and cost effectiveness from factorial designs come at the cost of not controlling size and an increased rate of false positives relative to a business-as-usual counterfactual. Alternatively, they come at the cost of a more complicated interpretation of the main results as a weighted-average of interactions with other treatments that may not represent a typical counterfactual. Further, using underpowered factorial designs to explore whether interactions are significant comes at the risk of overestimating the true effect, conditional on rejecting the null of no effect.

We also contribute to the literature that aims to improve the analysis of field experiments (e.g., Young, 2018; List et al., 2019). Our paper follows in this tradition by documenting a problem with the status quo, quantifying its importance, and identifying the most relevant recent advances in theoretical econometrics that can mitigate the problem. Specifically, we show that the econometric analysis of nonstandard inference problems can improve inference in factorial designs which are ubiquitous in field experiments.

Finally, we contribute to the literature on the pitfalls of focusing on statistical significance in applied work (e.g., Brodeur et al., 2016; Wasserstein & Lazar, 2016; Amrhein et al., 2019; Wasserstein et al., 2019; Brodeur et al., 2020). Specifically, the problems we highlight in this paper are less due to factorial designs per se. Rather they stem from the combination of a focus on statistical significance to assess if effects are meaningful, and most factorial experiments being under-powered to detect interactions.

II. Factorial Designs in Theory

A. Setup

This section discusses theoretical aspects of experiments with factorial (or "cross-cut") designs. We focus on factorial designs with two treatments, T_1 and T_2, ("2×2 designs"), where researchers randomly assign some subjects to receive treatment T_1, some subject to receive treatment T_2, and some subjects to receive both treatments (see table 1). The analysis straightforwardly extends to cross-cut designs with more than two treatments.

TABLE 1.—2×2 FACTORIAL DESIGN

		T_1	
		No	*Yes*
T_2	*No*	N_1	N_2
	Yes	N_3	N_4

N_j is the number of individuals randomly assigned to cell j.

We are interested in the causal effect of T_1 and T_2 on an outcome Y. We use the potential outcomes framework (Rubin, 1974). The potential outcomes $\{Y_{t_1,t_2}\}$ are indexed by both treatments, $T_1 = t_1$ and $T_2 = t_2$, and are related to the observed outcome as $Y = \sum_{t_1\in\{0,1\}} \sum_{t_2\in\{0,1\}} \mathbf{1}(T_1 = t_1, T_2 = t_2) \cdot Y_{t_1,t_2}$. We assume that both treatments are randomly assigned and independent of each other, which is common in practice (e.g., Olken, 2007; Bertrand et al., 2010).

B. Long and Short Regression Models

Researchers analyzing experiments based on cross-cut designs typically consider one of the following two population regression models:

Long (or fully saturated) model:

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 T_2 + \varepsilon \tag{1}$$

Short model:

$$Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s. \tag{2}$$

The long model (equation [1]) includes both treatment indicators as well as their interaction, while the short model (equation [2]) includes only the two treatment indicators.[6]

The population regression coefficients in the long regression model correspond to the main average treatment effects (ATEs) of T_1 and T_2 against a business-as-usual counterfactual (this counterfactual can also be interpreted as the outcomes in a pure experimental control group) and the interaction effect:

$$\beta_1 = E(Y_{1,0} - Y_{0,0}) \quad \text{(ATE of } T_1 \text{ relative to a counterfactual where } T_2 = 0\text{)}, \tag{3}$$

$$\beta_2 = E(Y_{0,1} - Y_{0,0}) \quad \text{(ATE of } T_2 \text{ relative to a counterfactual where } T_1 = 0\text{)}, \tag{4}$$

$$\beta_{12} = E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}) \quad \text{(interaction effect)}.^7 \tag{5}$$

By contrast, the regression coefficients in the short model are

$$\begin{aligned}\beta_1^s &= E(Y_{1,1} - Y_{0,1})P(T_2 = 1) \\ &\quad + E(Y_{1,0} - Y_{0,0})P(T_2 = 0) \qquad (6)\\ &= E(Y_{1,0} - Y_{0,0}) + E(Y_{1,1} - Y_{0,1} \\ &\quad - Y_{1,0} + Y_{0,0})P(T_2 = 1), \qquad (7)\end{aligned}$$

$$\begin{aligned}\beta_2^s &= E(Y_{1,1} - Y_{1,0})P(T_1 = 1) \\ &\quad + E(Y_{0,1} - Y_{0,0})P(T_1 = 0) \qquad (8)\\ &= E(Y_{0,1} - Y_{0,0}) + E(Y_{1,1} - Y_{0,1} \\ &\quad - Y_{1,0} + Y_{0,0})P(T_1 = 1). \qquad (9)\end{aligned}$$

Equation (6) shows that β_1^s yields a weighted average of the ATE of T_1 relative to a counterfactual where $T_2 = 1$ and the ATE of T_1 relative to a business-as-usual counterfactual where $T_2 = 0$. The weights, $P(T_2 = 1)$ and $P(T_2 = 0)$, are determined by the experimental design. Alternatively, β_1^s can be written as the sum of the ATE of T_1 relative to the $T_2 = 0$ counterfactual and the interaction effect multiplied by $P(T_2 = 1)$ (equation [7]). Equations (8) and (9) present the corresponding expressions for β_2^s. Unless the interaction effect is zero, β_1^s and β_2^s do not correspond to the main effects but yield composite treatment effects that are weighted averages of ATEs relative to different counterfactuals.

Remark 1. The problem of choosing between the long model and the short model is not unique to factorial designs and arises in many contexts. For example, when estimating treatment effects in observational studies, researchers need to decide whether to include the covariates linearly or consider fully interacted specifications (e.g., Angrist & Krueger, 1999; Angrist & Pischke, 2009). However, the practical implications are not the same because experimental treatments are fundamentally different in nature from standard covariates, as we discuss in section IIC. The choice between the short and the long model (with interactions between the treatment and strata indicators) is also relevant in stratified experiments (e.g., Imbens & Rubin, 2015; Ansel et al., 2018; Bugni et al., 2018, 2019).

C. Long or Short Model: What Do We Care about?

Section IIB shows that the short model yields a weighted average of treatment effects that depends on the nature and distribution of the other treatment arms in the experiment. This weighted average is typically neither of primary academic interest nor policy-relevant. This view is consistent with how papers we reanalyze motivate their object of interest, which is usually the main treatment effect against a business-as-usual counterfactual. Of the nineteen papers in table A1 in appendix A.1 that present results from the short model without all interactions, we did not find any study that mentioned (in the main text or a footnote) that the presented treatment effects should be interpreted as either (a) a composite effect that includes a weighted average of the interaction with the other treatments or (b) being against a counterfactual that was not business-as-usual but one that also had the other treatments in the same experiment.

One way to make the case for the short model is to recast the problem we identify as one of external rather than internal validity. Specifically, all experiments are carried out in a context with several unobserved "background" covariates. Thus, any experimental treatment effect is a weighted average of effects conditional on unobserved covariates. If the

[6] Following Angrist and Pischke (2009, chapter 3) and Hansen (2022, chapter 2), we interpret $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = E(XX')^{-1}E(XY)$, where $X = (1, T_1, T_2, T_{12})'$, as the population regression coefficient (or linear projection coefficient) and $\varepsilon = Y - X'\beta$ as the population residual (or projection error). Similarly, we interpret $\beta^s = (\beta_0^s, \beta_1^s, \beta_2^s)' = E(XX')^{-1}E(XY)$, where $X^s = (1, T_1, T_2)'$, and $\varepsilon^s = Y - X^s\beta^s$ as the population regression coefficient and the population residual, respectively.

[7] The interaction effect is the difference between the effect of jointly providing both treatments and the sum of the main effects.

other experimental arms are considered analogous to unobserved background covariates, inference on treatment effects based on the short model can be considered internally valid. In this view, the challenge is that the unobserved covariates (including other treatment arms) will vary across contexts.

However, experimental treatments are fundamentally different from standard background covariates. They are determined by the experimenter based on research interest and rarely represent real-world counterfactuals. In some cases, the interventions studied are new, and the other treatments may not even exist in the study population. Even if they do exist, there is no reason to believe that the distributions in the population mirror those in the experiment. Thus, we view this issue as a challenge to internal validity. Further, papers with factorial designs often use the two-step procedure described in section IIE and present results from the short model *after* mentioning that the interactions are not significantly different from zero (e.g., Banerjee et al., 2007; Karlan & List, 2007). This suggests that our view that interactions matter for internal validity is shared broadly.

Finally, even in settings where the coefficients in the short model are of interest, they can always be constructed based on the coefficients in the long model, but the converse is not true. One can also use the long model to test hypotheses about the coefficients in the short regression model: $H_0 : \beta_1^s = \beta_1 + \beta_{12}P(T_2 = 1) = 0$. Which test is more powerful depends on the relative sample size in the four experimental cells.[8] Unlike the short model, the long model additionally allows for testing a rich variety of hypotheses about counterfactual effects such as $H_0 : \beta_1 + \beta_{12}p = 0$ for policy-relevant values of p, which generally differ from the experimental assignment probability $P(T_2 = 1)$. For instance, résumé audit experiments may vary characteristics such as age, gender, race, education, and experience with the sample size allocated to various combinations of these characteristics being different from their proportion in the population. In such a case, short model estimates are difficult to interpret, whereas estimating the long model and calculating a weighted average of main and interaction effects with weights equal to their population proportions may yield a more policy-relevant treatment effect.

To summarize, the long model estimates all the underlying parameters of interest (the main effects and the interactions). In contrast, β_1^s is rarely of interest in its own right, and even if it is, the long model allows for estimation and inference on β_1^s as well.

D. Inference on Main Effects

Suppose that the researcher has access to a random sample $\{Y_i, T_{1i}, T_{2i}\}_{i=1}^N$. Consider the problem of testing hypotheses about the main effect of T_1 relative to a business-as-usual counterfactual: $H_0 : \beta_1 = E(Y_{1,0} - Y_{0,0}) = 0$.

To illustrate, suppose the data-generating process is given by

$$Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} T_{2i} + \varepsilon_i,$$
$$\varepsilon_i \sim N(0, \sigma^2), \tag{10}$$

where ε_i is independent of (T_{1i}, T_{2i}) and σ^2 is known. If the interaction effect β_{12} is zero, conditional on $\{T_{1i}, T_{2i}\}_{i=1}^N$, $\hat{\beta}_1 \sim N(\beta_1, \mathrm{Var}(\hat{\beta}_1))$ and $\hat{\beta}_1^s \sim N(\beta_1, \mathrm{Var}(\hat{\beta}_1^s))$, where $\mathrm{Var}(\hat{\beta}_1) = \sigma^2(\frac{1}{N_1} + \frac{1}{N_2}) \geq \mathrm{Var}(\hat{\beta}_1^s) = \sigma^2(\frac{N_1N_3+N_1N_4+N_2N_3+N_2N_4}{N_1N_2N_3+N_1N_2N_4+N_1N_3N_4+N_2N_3N_4})$. As a result, the short model t-test exhibits higher power than the long model t-test.

If, on the other hand, $\beta_{12} \neq 0$, ignoring the interaction can lead to substantial size distortions. To illustrate, we introduce a simple running example. Consider a 2×2 design with a total sample size of $N = 1{,}000$ and $N_1 = N_2 = N_3 = N_4 = 250$. The data are generated based on Model (10) with $\varepsilon_i \sim N(0, 1)$, T_{1i} and T_{2i} randomly assigned and independent of each other, and $P(T_{1i} = 1) = P(T_{2i} = 1) = 0.5$. This design has power 90% to detect an effect of 0.2σ (0.29σ) at the 5% level using the short model (long model).

Figure 1 shows how power, bias, and size vary across different values of β_{12} in both the long and the short model. When $\beta_{12} = 0$, the short model t-test controls size and exhibits higher power than the long model t-test as discussed before. However, these power gains come at the cost of bias and size distortions whenever $\beta_{12} \neq 0$. Importantly, even modest values of $|\beta_{12}|$ lead to considerable size distortions. For instance, $|\beta_{12}| > 0.1\sigma$ more than doubles the rate of false rejection of the null (in the data we reanalyze in section IIIB, we find that $|\hat{\beta}_{12}| > 0.1\sigma$ in over 36% of cases). By contrast, the long model is unbiased and exhibits correct size for all values β_{12}. The main takeaway from figure 1 is that researchers should avoid the short model for making inference on the main effects, unless they are certain that $\beta_{12} = 0$.

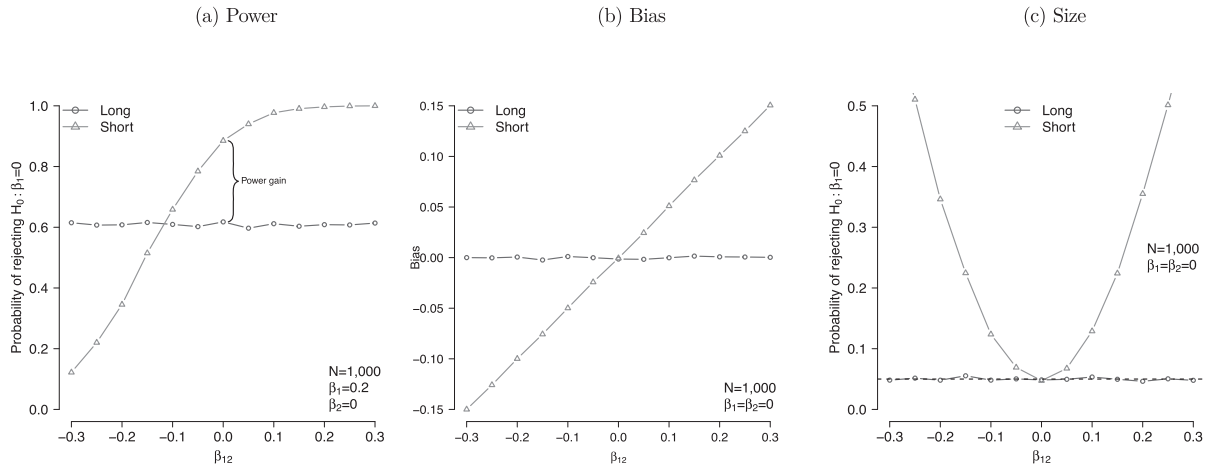
E. Model Selection (or Pretesting) Yields Invalid Inferences

Researchers often recognize that using the short model is only correct for inference on the main treatment effect if the interaction is close to zero (as implied by the quote from Kremer (2003) in the introduction). However, the problem is that the value of the interaction is unknown ex ante. Therefore, a common practice is to employ a data-driven two-step procedure to determine whether to ignore the interaction:

1. Estimate the long model and test the null hypothesis that β_{12} is zero (i.e., $H_0 : \beta_{12} = 0$) using a two-sided t-test.
2. (a) If $H_0 : \beta_{12} = 0$ is rejected, test $H_0 : \beta_1 = 0$ using the long model t-test.
   (b) If $H_0 : \beta_{12} = 0$ is not rejected, test $H_0 : \beta_1 = 0$ using the short model t-test.

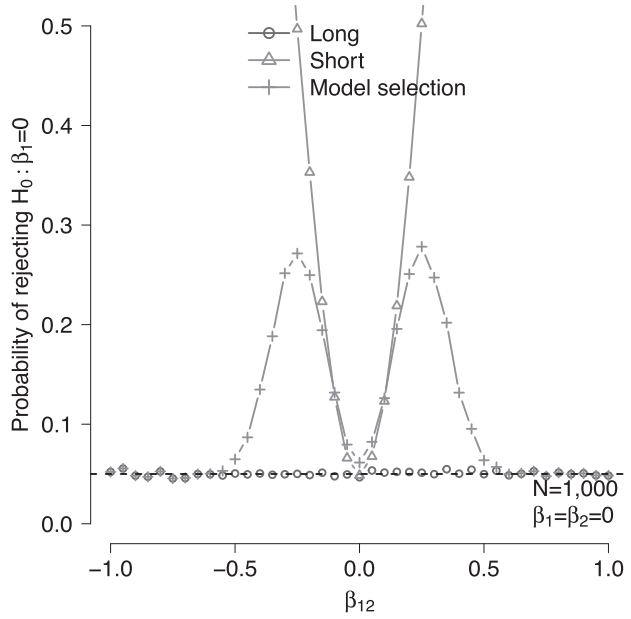
[8] In practice, we recommend comparing both tests when doing power calculations. If both tests have the same power, the short model is more straightforward.

FIGURE 1.—THE PERCEIVED POWER GAINS FROM THE SHORT MODEL COME AT THE COST OF BIASED ESTIMATORS AND NOT CONTROLLING SIZE UNLESS β_{12} IS EXACTLY EQUAL TO ZERO



Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size for figures 1a and 1c is $\alpha = 0.05$.

FIGURE 2.—MODEL SELECTION DOES NOT CONTROL SIZE



Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size is $\alpha = 0.05$. For the model selection, the short model is estimated if one fails to reject $\beta_{12} = 0$ at the 5% level.

While seemingly attractive, such data-dependent model selection leads to invalid inferences (e.g., Leeb & Pötscher, 2005, 2006, 2008; Kahan, 2013). Figure 2 shows the size properties of the two-step model selection approach in our running example. For reference, we also include results for the short and long model t-tests. The main takeaway from figure 2 is that model selection leads to incorrect inferences and false positives for a wide range of values of β_{12}.[9] Model selection can be particularly problematic for program evaluation field experiments because they are expensive to run, and therefore typically not adequately powered to reject that the interactions are zero (section IIF).

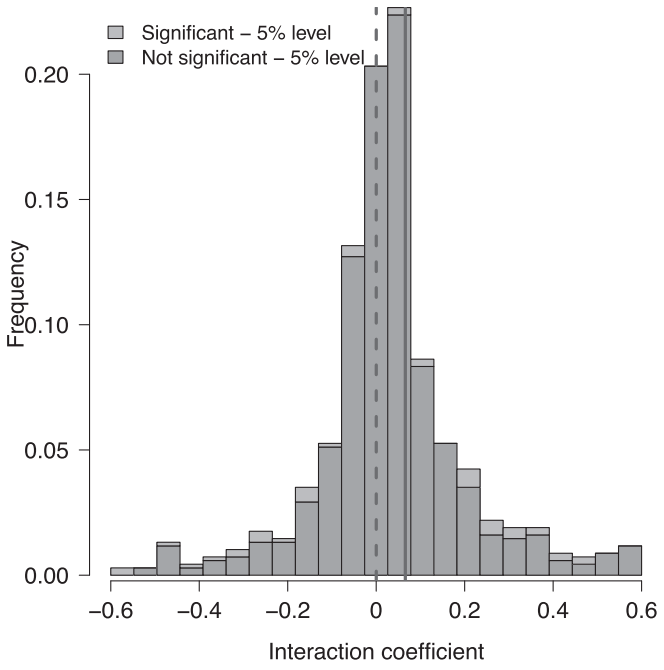
The range of values for $|\beta_{12}|$ for which model selection leads to substantial size distortions shrinks as the sample size (and power) of the experiment increases. However, it can be quite large in realistic settings. In our running example, with 1,000 observations one would need $|\beta_{12}|$ to be above 0.5 to avoid notable size distortions. Even with 10,000 observations, only values of $|\beta_{12}|$ above 0.2 lead to negligible size distortions (see figure A13). The true value of the interaction is unknown and likely to be in this "problematic range" in many practical settings (see figure 3), and so we recommend that researchers avoid the data-driven model-selection approach.

Remark 2. As figure 2 shows, model selection is less of a concern when the interactions are either zero or very large, but is a first-order issue when interactions are in the problematic range noted above. This issue is relevant in many settings. For instance, Banerjee et al. (2021) have proposed a LASSO-based method for selecting and making inferences on the most effective combination of treatments. However, they do so by imposing the restriction that "[treatments and their interactions] have either no effect or have sufficiently large (positive or negative) influence on the outcomes."[10] In other words, they avoid the problem noted above by assuming that the interactions are outside the "problematic range" in figure 2. While their goal differs from ours (making inferences on the best treatment combination versus making inferences on main and interaction effects), this example illustrates the continued prevalence of model selection in the analysis of field experiments.

[9] This is true even when $\beta_{12} = 0$ (as seen in the blue line with crosses in figure 2) because the tests in the first and second step are not independent.

[10] See their assumption 3 and footnote 11 for a formal statement.

FIGURE 3.—DISTRIBUTION OF THE ESTIMATED INTERACTION EFFECTS



This figure shows the distribution of the interactions between the main treatments ($N = 868$ in this figure). We trim the top and bottom 1% of the distribution. The median interaction for these papers is 0.00σ (dashed vertical line), the median absolute value of the interaction is 0.07σ (solid vertical line), and the median relative absolute value of the interaction with respect to the main treatment effect is 0.37. Here 6.2% of interactions are significant at the 10% level, 3.6% are significant at the 5% level, and 0.9% are significant at the 1% level.

F. Inference on Interaction Effects

An alternative motivation for factorial designs is to learn about interactions and jointly explore the parameter space of main and interaction effects.

However, detecting interaction effects requires much larger sample sizes than needed for detecting main effects. To illustrate, we compare the standard errors of the OLS estimator of the interaction effect, $\hat{\beta}_{12}$, and the main effect, $\hat{\beta}_1$. Under the assumptions in section IID, the standard errors are $SE(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$ and $SE(\hat{\beta}_{12}) = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4}}$. Since $SE(\hat{\beta}_1) < SE(\hat{\beta}_{12})$, the power for detecting interaction effects is always lower than the power for detecting main effects, and the required sample size for detecting interaction effects is always larger than the required sample size for detecting main effects of equal magnitude. For example, we need eight times the sample size to have the same power to detect an interaction effect as to detect the main effect, when the interaction is half the size of the main effect (see appendix A.3). Given the more onerous sample size requirements to detect interactions relative to main effects, it is not surprising that only a few of the interaction effects are significant in the reanalysis in section IIIB.

Further, even when interactions estimates are significant, they can be misleading because significant results in underpowered studies are much more likely to reflect an outlier estimate of the interaction. In particular, low power is associated with a high *Type-M error* (or *exaggeration ratio*) (Gelman & Carlin, 2014). The Type-M error is the expectation of the absolute value of the estimator in a hypothetical replication study based on the same design as the original study, *conditional* on being significant, divided by the true effect (see p. 643 and figure 1 in Gelman & Carlin, 2014). For example, if the experiment has 80% power to detect treatment effects of 0.2σ or larger at the 5% level using the long model and the true value of the interaction is 0.1σ, then the Type-M error for $\hat{\beta}_{12}$ is $\sim 251\%$. That is, the estimator of the interaction would, on average, be over two times larger than the true value, conditional on being significant. Figure A9 in appendix A.3 shows the relationship between the Type-M error and the power of the experiment.

Note that using the long model to estimate and learn about interactions is fine since the long model estimator is always consistent and asymptotically normal, even if noisy in finite samples. The problem we document here arises because of the focus on statistical significance to assess whether a result is meaningful. Combined with the well-documented publication bias toward significant results (e.g., Franco et al., 2014; Andrews & Kasy, 2018; Christensen & Miguel, 2018; Abadie, 2020), the discussion above suggests that published results from under-powered studies are likely to meaningfully exaggerate the true effect. Following Gelman and Carlin (2014), we suggest studies report power to detect interactions (as well as Type-M errors) in their preanalysis plan.

III. Factorial Designs in Practice

In this section, we document common practices among researchers studying field experiments with factorial designs.

A. Data and Descriptive Statistics

We analyze all articles published between 2007 and 2017 in the top-five journals in economics.[11] Of the 3,505 articles published in this period, 124 (3.5%) are field experiments (table A6 provides more details). Factorial designs are widely used: Among 124 field experiments 27 (22%) had a factorial design.[12] Only eight of these 27 articles with factorial designs (∼30%) used the long model including all interaction terms as their main specification (see table 2).

[11] These journals are *The American Economic Review*, *Econometrica*, *The Journal of Political Economy*, *The Quarterly Journal of Economics*, and *The Review of Economic Studies*. We exclude the May issue of *The American Economic Review*, known as "AER: Papers and Proceedings."

[12] We do not consider two-stage randomization designs as factorial designs. A two-stage randomization design is where some treatment is randomly assigned in one stage. In the second stage, treatment status is rerandomized to study behavioral changes conditional on a realization of the previous treatment. Examples of studies with two-stage randomization designs include Karlan and Zinman (2009), Ashraf et al. (2010), and Cohen and Dupas (2010). Finally, we do not include experiments where there is no "treatment," but rather conditions are randomized to elicit individuals preference parameters (e.g., Andersen et al., 2008; Fisman et al., 2008; Gneezy et al., 2009).

TABLE 2.—FIELD EXPERIMENTS PUBLISHED IN TOP-FIVE JOURNALS BETWEEN 2007 AND 2017

	AER	ECMA	JPE	QJE	ReStud	Total
Field experiments	43	9	14	45	13	124
With factorial designs	11	2	4	6	4	27
Interactions included	3	1	1	2	1	8
Interactions not included	8	1	3	4	3	19

B. Ignoring Interactions in Practice

In section IID, we have shown that ignoring interactions can lead to substantial size distortions and false positives. Here we examine the practical implications of ignoring the interactions in the papers listed in table A1. We reanalyze the data from all field experiments with factorial designs and publicly available data that do not include all the interactions in the main specification.[13] Of the ten most-cited papers with factorial designs listed in table A1, only one includes all the interactions in the main specification. More recent papers (which are less likely to be among the most cited) are more likely to include all interaction terms. Out of the 27 papers with factorial designs published in top-five journals, 19 papers do not include all interaction terms (over 70%).[14] Of these 19, four papers did not have publicly available replication data. In an online appendix we describe the experimental design of each of the 27 papers and provide details on our replication analysis.[15]

We downloaded the publicly available data files and replicated the main results in each of the remaining fifteen papers. We standardized the outcome variable in each paper to have mean zero and standard deviation of one. We then compared the original treatment effects (estimated without the interaction terms) with those estimated including the interaction terms.[16] In other words, we compare estimates based on the short model (equation [2]) to those based on the long model (equation [1]).

Key facts about interactions. As the discussion in section IID highlights, the extent to which the short model will not control size depends on the value of the interactions in practice. We therefore start by plotting the distribution of estimated interaction effects (figure 3) and documenting facts regarding interactions from our reanalysis. We find that interactions are quantitatively important and typically not second order. All estimates are measured in standard deviations (σ) of the outcome variable. Although the median (mean) interaction for these papers is 0.00σ (0.00σ), the median (mean) *absolute* value of the interaction is 0.07σ (0.13σ). The median (mean) absolute value of interactions relative to the main treatment effects is 0.37 (1.55). Thus, although it may be true that interactions are small on average across all studies, they are often sizeable in any given study. In our data, the absolute value of the interactions is greater than 0.1σ in 36% and greater than 0.2σ in 19% of the cases. These magnitudes lead to a 12% and 35% chance of rejecting the null of no effect in our running example (as seen in figure 1), which corresponds to more than a doubling and a sextupling, respectively, in the rate of false rejections at the 5% level.

The second key finding is that most experiments will rarely reject the null hypothesis that the interactions are zero (figure 3 shades the fraction of the interactions that are significant in the studies that we reanalyze). Among the fifteen papers that we reanalyzed, 6.2% of interactions (spread across four papers) are significant at the 10% level, 3.6% are significant at the 5% level (spread across three papers), and 0.9% are significant at the 1% level (in one paper).[17] These findings are not surprising because factorial designs are rarely powered to detect meaningful interactions.

The fact that most experiments were not explicitly powered to detect interactions suggests that the main reason for running experiments with factorial designs seems to be the increase in power for detecting main effects. However, as we show below, this comes at the considerable cost of an increased rate of false positives (which is unsurprising based on the distribution of interactions shown in figure 3).

Ignoring interactions has important implications for estimation and inference. Figure 4a compares the original treatment effect estimates based on the short model to the estimates based on the long model which includes the interaction terms (figure 4b zooms in to cases where the value of the main treatment effects in the short model is between -1 and 1 standard deviation). The median change in the absolute value of the point estimate of the main treatment effect is 96%. Roughly 26% of estimated treatment effects change sign when they are estimated using the long regression.

Table 3 shows how the significance of the main treatment estimates changes when using the long instead of the short model. About 48% of treatment estimates that were significant at the 10% level based on the short model are no longer significant based on the long model. Here 53% and 57% of estimates lose significance at the 5% and 1% levels, respectively. A much smaller fraction of treatment effects that were not significant in the short model are significant based on the long regression (6%, 5%, and 1%, at the 10%, 5%, and 1% levels, respectively).[18]

[13] We also reanalyze the effect of not including the interaction in the studies that do include all the interactions in their main specification in appendix A.1.4.

[14] Although we restrict our reanalysis to papers published in "top-five" journals, factorial designs are also prevalent in papers published in lower-ranked journals. Hence, the total number of articles focusing on the short model published in this period is likely much larger.

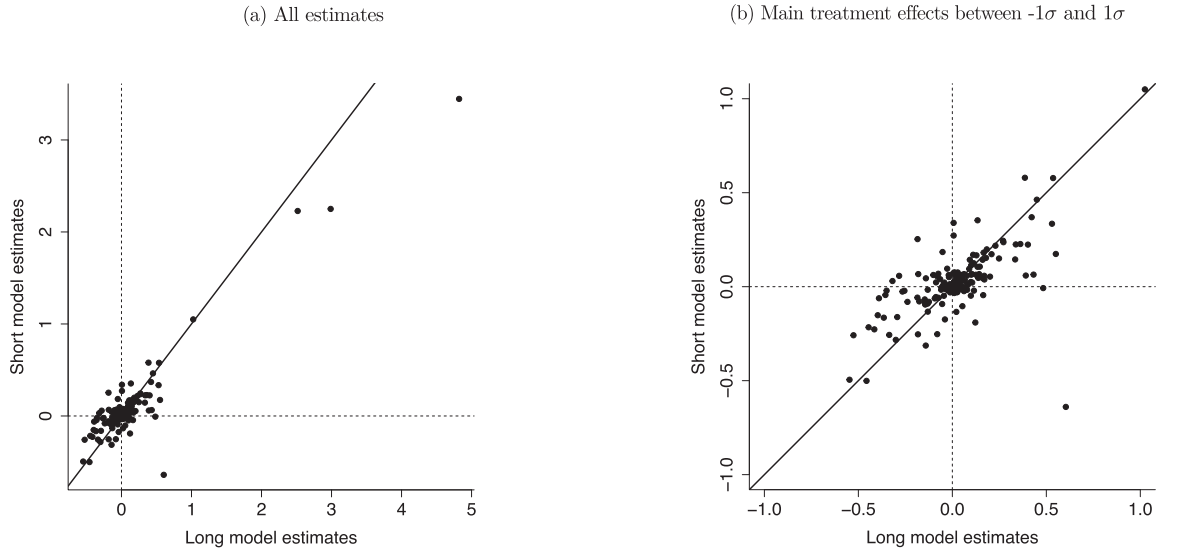
[15] Available at http://mauricio-romero.com/pdfs/papers/Appendix_cross cuts.pdf.

[16] If studies have factorial designs that cross-randomize more than two treatments, we include only two-way interactions in this calculation.

[17] Among the papers that originally included all interactions, 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0.0% are significant at the 1% level. See appendix A.1.4 for more details.

[18] These results are not driven by just a few papers. If we first estimate the median change in the absolute value of the estimate *within* each paper, and

FIGURE 4.—TREATMENT EFFECTS ESTIMATES BASED ON THE LONG AND THE SHORT MODEL



This figure shows how the main treatment estimates change between the short and the long model across all studies ($N = 172$ in this figure). Figure 4a has all the treatment effects, and figure 4b zooms in to cases where the value of the main treatment effects in the short model is between -1 and 1 standard deviation. The median main treatment estimate from the short model is 0.01σ, the median main treatment estimate from the long model is 0.02σ, the average absolute difference between the treatment estimates of the short and the long model is 0.05σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 96%, and 26% of treatment estimates change sign when they are estimated using the long model instead of the short model.

TABLE 3.—SIGNIFICANCE OF TREATMENT EFFECTS ESTIMATES BASED ON THE LONG AND THE SHORT MODEL

Panel A: Significance at the 10% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	95	34	129
Significant	6	37	43
Total	101	71	172
Panel B: Significance at the 5% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	111	29	140
Significant	6	26	32
Total	117	55	172
Panel C: Significance at the 1% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	140	17	157
Significant	2	13	15
Total	142	30	172

This table shows the number of significant coefficients at a given level when estimating the long regression (columns) and the short regression (rows). It includes information from all papers with factorial designs and publicly available data that do not include the interactions in the original study. Panel A uses a 10% significance level, panel B uses 5%, and panel C uses 1%.

then the median change across papers, the result is similar to estimating the median absolute changes across all estimates at 97%. Likewise, if we first estimate the proportion of estimates that change sign within each paper, and then estimate the average across papers, the result is 25%, which is similar to estimating the proportion of estimates that change sign. Finally, 73% of papers have at least one estimate that is no longer significant at the 10% level when estimating the full model, 77% have at least one estimate that is no longer significant at the 5% level, and 82% have at least one estimate that is no longer significant at the 1% level.

We find similar results when we restrict our reanalysis to the ten most cited papers with factorial designs that do not include the interaction terms (with data available for reanalysis). When we reestimate the treatment effects in these papers after including interactions, we find that out of 21 results that were significant at the 5% level in the paper, nine (or 43%) are no longer so after including interactions. Corresponding figures and tables are presented in appendix A.1.2 (figure A2 and table A2).

Finally, we also distinguish between policy and conceptual experiments in table A1 (the latter typically have more treatments and interactions) and see that the problem of incorrect inference from ignoring interaction terms remains even when we restrict attention to the policy experiments. Of the twelve policy experiments, nine do not include all interactions. When we reestimate the treatment effects in these nine papers after including interactions, we find that out of nineteen results that were significant at the 5% level in the paper, six (or 32%) are no longer so after including interactions. Corresponding figures and tables are presented in appendix A.1.3 (figure A4 and table A3).[19]

IV. Improving Power for Detecting Main Effects

We now examine whether it is possible to improve power for detecting main effects relative to long model t-tests while maintaining size control for relevant values of the interactions. We consider 2×2 factorial designs and briefly

[19] Among the papers that originally included all interactions, 23% of results that are significant at the 5% level in the short model are not significant in the long model. See appendix A.1.4 for more details.

comment on factorial designs with more than two treatments at the end of each subsection. Throughout we will focus on the main ideas underlying the different econometric methods. Appendix A.4 provides detailed descriptions and implementation details.

A. *Setup*

We focus on β_1 and partial out T_2 and the constant, keeping the partialing out implicit. Defining $T_{12} = T_1 T_2$, the regression model of interest is

$$Y = \beta_1 T_1 + \beta_{12} T_{12} + \varepsilon. \tag{11}$$

Our goal is to test hypotheses about the main effect β_1.

The two-sided long model t-test is the uniformly most powerful test among tests that are unbiased for all values of the interaction effect (e.g., van der Vaart, 1998; Elliott et al., 2015a).[20] This implies that any test that is more powerful than the long model t-test for some values of β_{12} must have lower power somewhere else. Thus, to achieve higher power than the long model t-test, one has to choose which values of β_{12} to direct power to based on prior knowledge.

If one insists on size control for all β_{12}, the scope for power improvements relative to the long model t-test is theoretically limited.[21] For example, at the 5% level, the maximal theoretically possible power improvement over the long model two-sided t-test is 12.5 percentage points. Section IVB proposes a nearly optimal test that comes close to achieving the maximal power gain at a priori likely values of the interaction, while controlling size for all values of the interaction. In appendix A.6, we show that a Bonferroni-style correction after model selection leads to local power improvements for a range of positive values of the interaction.

The limited scope for power improvements relative to the long model t-test motivates relaxing the uniform size control requirement and imposing additional restrictions on β_{12}. An extreme example is the short model t-test, which can improve power relative to long model t-test by much more than 12.5%, but controls size only under the restrictive assumption that $\beta_{12} = 0$. In section IVC, we explore an intermediate approach that restricts the magnitude of β_{12}, which is often more realistic than assuming that β_{12} is exactly equal to zero.

B. *Nearly Optimal Tests Targeting Power toward a Likely Value* $\bar{\beta}_{12}$

Suppose that a particular value $\beta_{12} = \bar{\beta}_{12}$ is a priori likely and that we want to find a test that controls size for all values of β_{12} and is as powerful as possible when $\beta_{12} = \bar{\beta}_{12}$. For concreteness, we focus on the case where $\bar{\beta}_{12} = 0$ and consider the testing problem

$$H_0 : \beta_1 = 0,\ \beta_{12} \in \mathbb{R} \quad \text{against} \quad H_1 : \beta_1 \neq 0,\ \beta_{12} = 0. \tag{12}$$

We use the numerical algorithm developed by Elliott et al. (2015a,b) to construct a nearly optimal test for the testing problem in equation (12).[22] Elliott et al. (2015a) consider a setting where one is interested in maximizing weighted average power. The best test in this setting is a Neyman-Pearson test based on the least favorable distribution (LFD). The LFD is often difficult to compute analytically, and so Elliott et al. (2015a) instead focus on an approximate LFD, which yields feasible and nearly optimal tests.

Figure 5 displays the results of applying the nearly optimal test in our running example. The test controls size for all values of β_{12} and, by construction, is nearly optimal when $\beta_{12} = 0$. For example, when $\beta_1 = 0.2$ the power of the nearly optimal test is 98.5% of the maximal possible power at $\beta_{12} = 0$ (implied by the corresponding uniformly most powerful one-sided t-test). A comparison with the long model t-test shows that the nearly optimal test is more powerful when β_{12} is close to zero.

However, these power gains come at a cost. For certain values of β_{12}, the power can be much lower than that of the long model t-test. Appendix A.7.3 provides a comprehensive assessment of the performance of the nearly optimal tests by plotting power curves for different values of β_1.

Finally, the nearly optimal test of Elliott et al. (2015a) becomes computationally prohibitive with many interactions (i.e., many nuisance parameters) and, thus, cannot be recommended for complicated factorial designs. The Bonferroni approach of McCloskey (2017, 2020) discussed in appendix A.6 constitutes a possible alternative in such settings.

C. *Inference under a Priori Restrictions on the Magnitude of* β_{12}

If the researcher is certain that $\beta_{12} = \bar{\beta}_{12}$, they can obtain powerful tests based on a regression of $Y - \bar{\beta}_{12} T_{12}$ on T_1. If $\bar{\beta}_{12} = 0$, this corresponds to the short model t-test. As shown in section IID, short model t-tests are more powerful than long model t-tests when $\beta_{12} = 0$, but do not control size when $\beta_{12} \neq 0$.

Exact knowledge of β_{12} may be too strong of an assumption. Suppose instead that the researcher imposes prior knowledge in the form of a restriction on the magnitude of the interaction effect β_{12}.

Assumption 1. $|\beta_{12}| \leq C$ *for some* $C < \infty$.

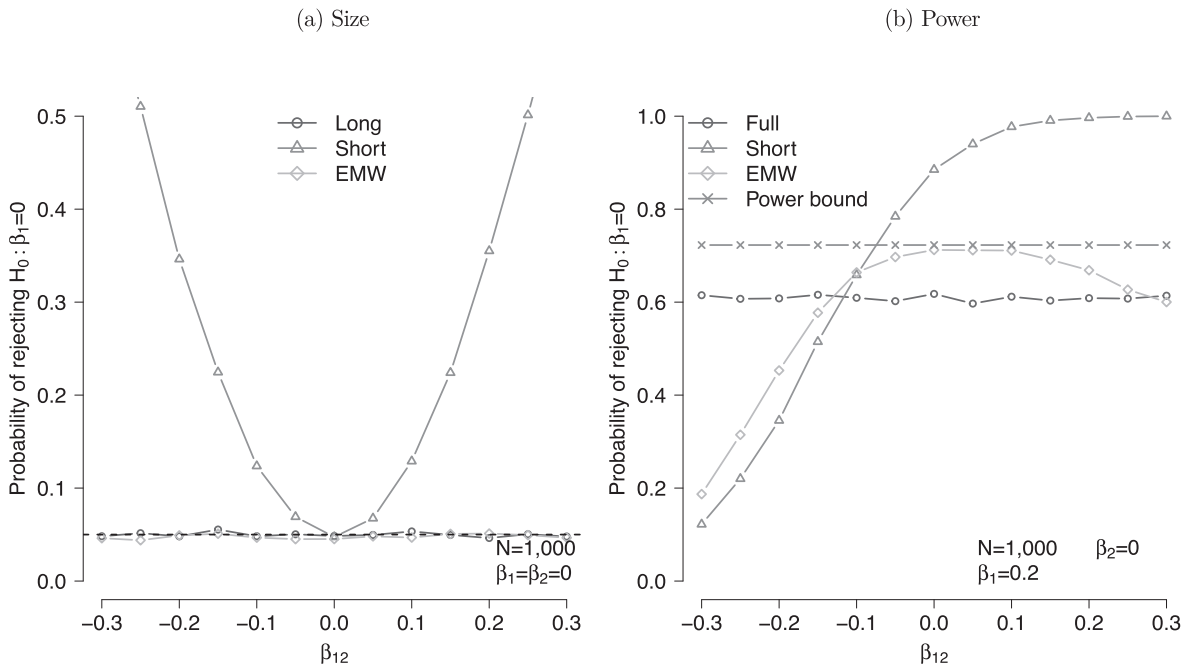
Assumption 1 restricts the parameter space for β_{12} and implies that $\beta_{12} \in [-C, C]$. We explore two different approaches for making inferences under this assumption. First,

[20] A test is unbiased if its power is larger than its size.

[21] This is because the one-sided long model t-tests are uniformly most powerful (e.g., Proposition 15.2 in van der Vaart, 1998) so that, for any β_{12}, the maximal power is achieved by a one-sided t-test (e.g., Armstrong & Kolesar, 2015, 2021). See Armstrong and Kolesar (2018) for a discussion of the implications for confidence intervals.

[22] Our code to implement this procedure for 2×2 factorial designs is available at https://mtromero.shinyapps.io/elliott/.

FIGURE 5.—THE NEARLY OPTIMAL TEST OF ELLIOTT ET AL. (2015A) CONTROLS SIZE AND YIELDS POWER GAINS OVER RUNNING THE FULL MODEL NEAR $\tilde{\beta}_{12} = 0$



Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size for figures 5a and 5b is $\alpha = 0.05$. EMW refers to the nearly optimal test of Elliott et al. (2015a). The power bound in figure 5b is the power of the one-sided long model t-test for the testing problem $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 > 0$.

we construct optimal confidence intervals under assumption 1 based on the approach developed by Armstrong et al. (2020). Their confidence intervals are based on linear estimators for β_1 and account for the worst case bias of the estimators. As a result, the length of the confidence interval is determined by the bias and the variance of the estimator, and to obtain optimal confidence intervals one has to solve a bias-variance trade-off. This problem can be solved using convex optimization. We refer to this approach as the Armstrong-Kolesar-Kwon (AKK) approach.

The second approach is based on constructing bounds on the main effect implied by assumption 1. In particular, upper and lower bounds on β_1 can be obtained from regressions of $Y + CT_{12}$ on T_1 and $Y - CT_{12}$ on T_1, respectively. We apply the procedure of Imbens and Manski (2004) and Stoye (2009) to construct valid confidence intervals for β_1. We refer to this approach as the Imbens-Manski-Stoye (IMS) approach.[23]

In figure 6, we report the rejection probabilities of tests that reject if zero is not in the AKK and IMS confidence intervals. To illustrate, we assume that $C = 0.1$, implying that $\beta_{12} \in [-0.1, 0.1]$.[24] Our results suggest that AKK and IMS can be substantially more powerful than long model t-tests when the prior knowledge is correct, but may exhibit size distortions when it is not. Panel b shows that the AKK and IMS power curves cross at zero. Thus, the choice between the two approaches should be based on which values of the interaction the researchers want to direct power to. Appendices A.7.4 and A.7.5 present the corresponding power curves for different values of β_1.

When researchers are primarily interested in the main effects and feel confident that the interactions are second order, AKK and IMS should be strictly preferred to the short model, since it is more realistic to prespecify that the interaction is in a range than exactly zero. However, prespecifying the appropriate range of prior values for the interaction is nontrivial and requires judgment.[25]

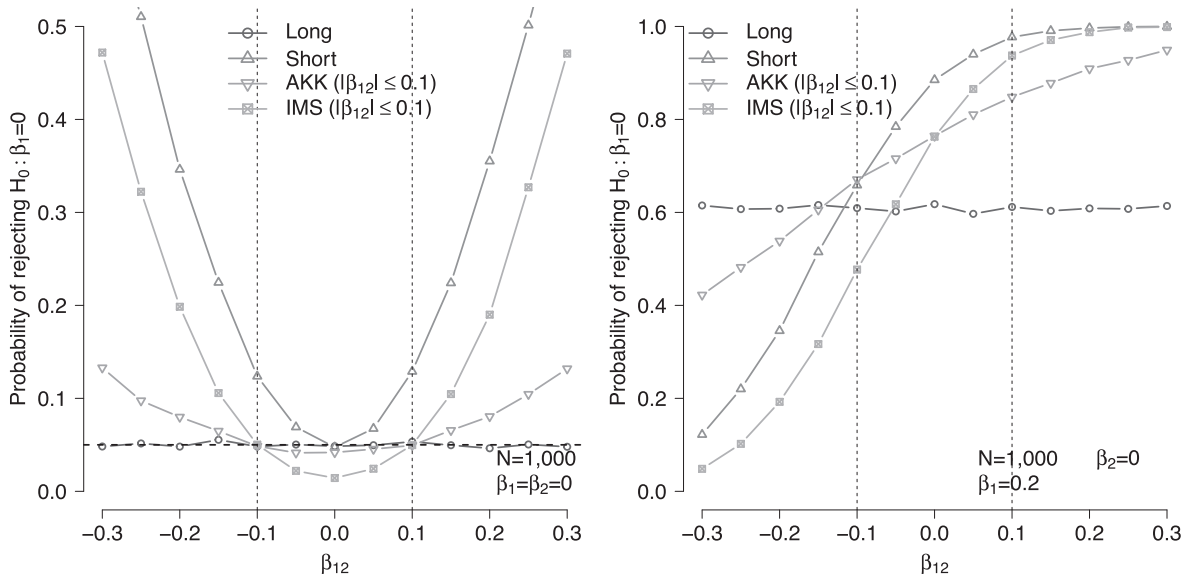
[23] As outlined in appendix A.4.3, it is straightforward to use the IMS approach if the prior information takes the form $C_1 \leq \beta_{12} \leq C_2$ for any $-\infty < C_1 < C_2 < \infty$, which may be more appropriate in some settings. Further, one could make inferences under restrictions on the direction of the interaction effects using the approach by Ketz and McCloskey (2023). Both types of approaches may be suitable in cases where there is a strong prior that treatments are complements or substitutes.

[24] Note that in our simulations $\sigma = 1$. This is similar to standardizing the outcome by the sample variance in the control group. Thus, the scale of the coefficients (β_1, β_2, and β_{12}) and of C can be interpreted as "standard deviations of the outcome." As mentioned above, in the papers we replicate, the median (mean) *absolute* value of the interaction is 0.07 (0.13) of the standard deviation of the outcome. Further, the absolute value of the interactions is greater than 10% of the standard deviation of the outcome in 36% of cases. Thus, in many settings it might be reasonable to assume $\beta_{12} \in [-0.1, 0.1]$, but researchers will need to judge, depending on the context, what a reasonable value for C is.

[25] It is problematic to use AKK or IMS based on first running the long model and not rejecting that the interaction is in a certain range. This would result in data-dependent model selection issue similar to those documented in section IIE. Thus, although AKK and IMS are improvements over the short model, they do not solve the underlying problem of not knowing the true value of the interaction.

FIGURE 6.—RESTRICTIONS ON THE MAGNITUDE OF β_{12} YIELD POWER GAINS IF THEY ARE CORRECT BUT LEAD TO INCORRECT INFERENCES IF THEY ARE NOT

(a) Size (b) Power



Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size for figures 6a and 6b is $\alpha = 0.05$. AKK refers to Armstrong et al. (2020)'s approach for constructing optimal confidence intervals under prior knowledge about the magnitude of β_{12}, $|\beta_{12}| \leq 0.1$ (dashed vertical lines). IMS refers to the Imbens and Manski (2004) and Stoye (2009) approach for constructing valid confidence intervals under prior knowledge about the magnitude of β_{12}, $|\beta_{12}| \leq 0.1$ (dashed vertical lines).

AKK and IMS remain computationally feasible in more complicated factorial designs. However, both approaches require reliable prior knowledge on the magnitude of potentially very many interactions to yield notable power improvements.

D. A Design-Based Approach for Improving Power

The discussion above focused on improving power for detecting main effects in existing experiments with factorial designs. Although these techniques can also be used to analyze new experiments (and be included in a preanalysis plan), a design-based alternative is to leave the "interaction cell" empty (i.e., to set $N_4 = 0$) and to reassign those subjects to the other cells (see table A5).

Leaving the interaction cell empty yields power improvements for testing hypotheses about the main effects relative to long model t-tests (see appendix A.5). Figure 7 provides an illustration based on our running example. Leaving the interaction cell empty yields tests that control size for all values of the interaction and achieve the highest power among the approaches with uniform size control (the long model t-test and the nearly optimal test).

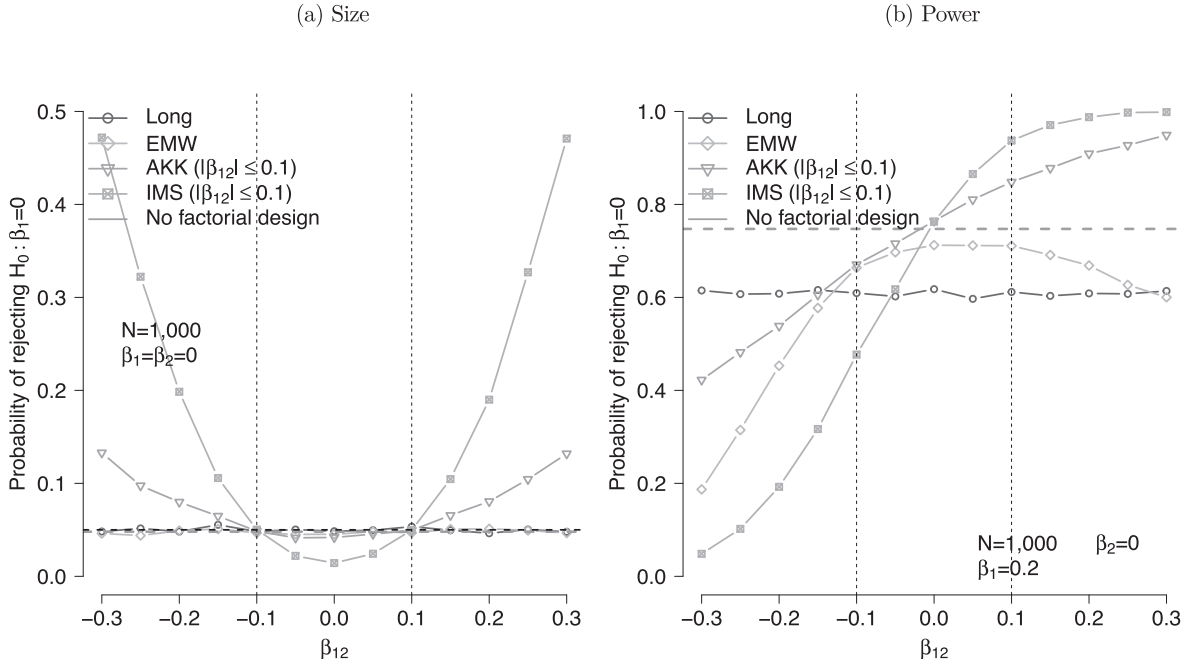
This design (with interaction cells empty) yields power gains relative to running two separate experiments because the control group is used twice, but it avoids the problem of interactions discussed above. An example of such a design is provided by Muralidharan and Sundararaman (2011) who study the impact of four different interventions in one experiment with one common control group, but no cross-cutting treatment arms.

E. Which Econometric Approach Should One Use in Practice?

For the design of new experiments, if the primary objects of interest are the main effects, we recommend leaving the interaction cells empty and increasing the number of units assigned exclusively to the treatment or the control groups. This design-based approach controls size and yields notable power improvements over the long model t-tests based on a factorial design.

For the reanalysis of existing experiments, the choice of the econometric method for making inferences on the main effects should be based on the strength of the available prior knowledge. If researchers have little prior knowledge about the interaction effects, we recommend using the long model t-tests, which are the uniformly most powerful unbiased tests. If prior knowledge about the interaction effects is available, but the researchers are not confident enough to be willing to sacrifice size control for all values of the interactions, we recommend the nearly optimal tests of Elliott et al. (2015a). The nearly optimal test allows for targeting power based on prior knowledge while ensuring uniform size control. If precise prior knowledge about the interaction effects is available, researchers can use the AKK or the IMS approach to leverage such prior knowledge to improve power substantially. However, unlike the other methods, these two

FIGURE 7.—LEAVING THE INTERACTION CELL EMPTY INCREASES POWER RELATIVE TO APPROACHES THAT CONTROL SIZE FOR ALL β_{12}



Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size for figures 7a and 7b is $\alpha = 0.05$. EMW refers to Elliott et al. (2015a)'s nearly optimal test. AKK refers to Armstrong et al. (2020)'s approach for constructing optimal confidence intervals under prior knowledge about the magnitude of β_{12}. IMS refers to the Imbens and Manski (2004) and Stoye (2009) approach for constructing valid confidence intervals under prior knowledge about the magnitude of β_{12}. The design of the experiment with the empty interaction cell is optimal for achieving equal power to detect both main effects; see appendix A.5 for details.

approaches exhibit size distortions when the prior knowledge is incorrect.

Irrespective of which method researchers use to improve power by incorporating prior knowledge, such prior knowledge should be prespecified in the preanalysis plan. In addition, we recommend always complementing the results with long model t-tests (even if only in an appendix). These tests have desirable optimality properties and allow for communicating results without subjective priors about interactions.

In some high-dimensional factorial designs, estimating the long model with all interactions may not be realistic. In this case we recommend that the authors prespecify which interactions they will ignore and which treatments they will pool in the preanalysis plan. To avoid model selection issues, it is crucial that such choices are made ex ante (and prespecified) and not be data-driven.

V. When Does the Short Model Make Sense?

Our discussion so far shows how using factorial designs and ignoring interactions can lead to incorrect inferences relative to a business-as-usual counterfactual (or pure experimental control group). At the same time, this approach is widely used in practice, perhaps reflecting a perception that classic texts on experimental design endorse it. We revisit these texts and review the historical use of factorial designs in field experiments to clarify the conditions and caveats under which factorial designs and the short model may be appropriate. We highlight four relevant cases below.

The first case is where the goal of initial experiments is to explore several treatment dimensions in an efficient way to generate promising interventions for further testing. For example, Cochran and Cox (1957, p. 152) recommend factorial designs for "exploratory work where the object is to determine quickly the effects of a number of factors over a specified range." Examples of such experiments include (a) agricultural experiments that vary soil, moisture, temperature, fertilizer, and several other inputs and (b) online A/B testing where large technology companies run thousands of randomized experiments each year to optimize profits over several dimensions (e.g., Kohavi et al., 2020). Both sets of examples feature sequential testing, making factorial designs an efficient way to quickly learn about which of several treatment dimensions that could be manipulated may be worth studying and testing further. In contrast, policy experiments are typically run only once, making factorial designs and short model estimates less desirable.

The second case is when the goal of the experiment is not hypothesis testing but to minimize MSE criteria (or other loss functions), which involve a bias-variance trade-off in estimating the main effects. For example, for small values of the interaction effects, estimators based on the short model can yield a lower root MSE than the estimators based on the design, which leaves the interaction cell empty (Blair et al., 2019). These alternative criteria also justify the use of factorial designs for agricultural experiments and online A/B testing, because their goal is to optimize decision making over

several factors (to maximize yields or profits) as opposed to testing if individual factors are "significant." Again, this contrasts with the case of policy experiments, where the goal is typically to test if a program or policy had a significant effect, and factorial designs and short-model inferences may therefore be problematic.

The third case is to improve an experiment's external validity. Cochran and Cox (1957, p. 152) recommend factorial designs for "experiments designed to lead to recommendations that must apply over a wide range of conditions. Subsidiary factors may be brought into an experiment so as to test the principal factors under a variety of conditions similar to those that will be encountered in the population to which recommendations are to apply"; see also the discussion in Fisher (1992). Thus, factorial designs and the short model may be fine when one dimension of the experiment is studying reasonable variants of the main treatment, but less so when all treatments are of primary interest.[26]

The fourth case is conceptual (as opposed to policy) experiments, such as résumé audit studies, where many or all of the characteristics that are randomized (e.g., age, education, race, and gender) do exist in the population. In these cases, a weighted average short model effect may be a reasonable target parameter subject to researchers indicating how the resulting effect should be interpreted. However, even for such experiments, we recommend (when feasible) designing the experiments such that the treatment share of various characteristics being studied is the same as their population proportion. Doing so will make the short-model coefficient more likely to approximate a population relevant parameter of interest.

VI. Conclusion

In this paper we study the theory and practice of inference in randomized experiments with factorial designs. These designs have been widely used and motivated by two main considerations: (i) studying more treatments in a cost-effective way and (ii) learning about interactions. We show that both of these uses can be problematic in practice, driven to a large extent by the lack of power to detect interactions.

Given our discussion and results, we recommend that (if realistic) studies using factorial designs should always present the fully saturated long regression model (even if only in an appendix) for transparency. If researchers would like to focus on results from the short model, they should clearly indicate that treatment effects should be interpreted as a composite effect that includes a weighted-average of interactions with other treatments. Further, if the estimand of interest is based on the short model, this should be specified in a preanalysis plan, and not justified ex post based on estimated interactions being insignificant (due to the problem of data-dependent model selection).

In practice, researchers' use of factorial designs and the short model is often motivated by prior beliefs that the absolute values of the interactions are "small." In such cases, the econometric approaches we discuss allow power gains for inference against a business-as-usual counterfactual (over the long model) while maintaining size control for relevant values of the interaction. In such cases, we recommend that researchers prespecify their priors and intended econometric approach for inference.

If the primary objects of interest are the main effects, an alternative design is to leave the interaction cells empty. This design-based approach naturally controls size and yields notable power improvements. If interaction effects are of primary interest, we recommend that experiments be explicitly powered to detect interactions and to indicate this in the preanalysis plan (as, e.g., in Mbiti et al., 2019).

Recently, our recommendations have been characterized as too conservative by Banerjee et al. (2021), who propose a LASSO-based method for making inferences on the most effective combination of treatments. Applying their approach to high-dimensional factorial designs is appealing: it allows researchers to explore the parameter space of main and interaction effects. However, their method relies on the strong assumption that "[treatments and interactions] have either no effect or have sufficiently large (positive or negative) influence on the outcomes." This restriction avoids model selection issues by assumption. It may be a good approximation in highly powered experiments or when researchers have strong prior knowledge about effect sizes.

Finally, it is worth noting that factorial designs *do* provide an efficient way of learning about multiple treatments as well as their interactions in the same experiment. The problems we highlight stem in large part from using factorial designs *in conjunction with* a focus on statistical significance for inference on whether treatment effects or interactions are meaningful. This approach reflects the default frequentist paradigm in experimental economics. Going forward, Bayesian methods (that do not privilege a binary "significant or not" threshold for inference) may constitute a promising framework for efficient learning in experiments with cross-cutting designs (e.g., Kassler et al., 2019).

[26] For example, in Alatas et al. (2012), the primary treatment effect of interest is the impact of community-based targeting, but they also randomize different aspects of how to run the community meeting (which are reasonable variants of the main treatment).

REFERENCES

Abadie, Alberto, "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights* 2:2 (2020), 193–208. 10.1257/aeri.20190252

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review* 102:4 (2012), 1206–1240. 10.1257/aer.102.4.1206

Amrhein, Valentin, Sander Greenland, and Blake McShane, "Scientists Rise Up against Statistical Significance," *Nature* 567 (2019), 305–307. 10.1038/d41586-019-00857-9

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. E. Rutström, "Eliciting Risk and Time Preferences," *Econometrica* 76:3 (2008), 583–618. 10.1111/j.1468-0262.2008.00848.x

Andrews, Isaiah, and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review* 109 (2018), 2766–2794. 10.1257/aer.20180310

Angrist, Joshua D., and Alan B. Krueger, "Chapter 23—Empirical Strategies in Labor Economics" (pp. 1277–1366), in Orley C. Ashenfelter and David Card (eds.), Vol. 3 of *Handbook of Labor Economics* (Amsterdam: Elsevier, 1999). 10.1016/S1573-4463(99)03004-7

Angrist, Joshua D., and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton, NJ: Princeton University Press, 2009).

Ansel, Jason, Han Hong, and Jessie Li, "OLS and 2SLS in Randomized and Conditionally Randomized Experiments," *Jahrbücher für Nationalökonomie und Statistik* 238:3–4 (2018), 243–293.

Armstrong, Timothy B., and Michal Kolesar, "Optimal Inference in a Class of Regression Models" (2015), arXiv:1511.06028v2.

——— "Optimal Inference in a Class of Regression Models," *Econometrica* 86:2 (2018), 655–683. 10.3982/ECTA14434

——— "Sensitivity Analysis Using Approximate Moment Condition Models," *Quantitative Economics* 12:1 (2021), 77–108. 10.3982/QE1609

Armstrong, Timothy B., Michal Kolesar, and Soonwoo Kwon, "Bias-Aware Inference in Regularized Regression Models" (2020), arXiv:2012.14823.

Ashraf, Nava, James Berry, and Jesse M. Shapiro, "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review* 100:5 (2010), 2383–2413. 10.1257/aer.100.5.2383

Athey, Susan, and Guido W. Imbens, "The Econometrics of Randomized Experiments" (pp. 73–140), in Abhijit Vinayak Banerjee and Esther Duflo (eds.), *Handbook of Economic Field Experiments*, Vol. 1 (Amsterdam: Elsevier, 2017). 10.1016/bs.hefe.2016.10.003

Banerjee, Abhijit, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, Francine N. Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha, "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization," NBER working paper 28726 (2021).

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics* 122:3 (2007), 1235–1264. 10.1162/qjec.122.3.1235

Banerjee, Abhijit, and Esther Duflo, "Chapter 7 Growth Theory through the Lens of Development Economics" (pp. 473–552), in Philippe Aghion and Steven N. Durlauf (eds.), Vol. 1 of *Handbook of Economic Growth* (Elsevier, 2005). 10.1016/S1574-0684(05)01007-5

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman, "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *Quarterly Journal of Economics* 125:1 (2010), 263–306. 10.1162/qjec.2010.125.1.263

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys, "Declaring and Diagnosing Research Designs," *American Political Science Review* 113:3 (2019), 838–859. 10.1017/S0003055419000194

Brodeur, Abel, Nikolai Cook, and Anthony Heyes, "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review* 110:11 (2020), 3634–3660. 10.1257/aer.20190687

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8:1 (2016), 1–32. 10.1257/app.20150044

Bruhn, Miriam, and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1:4 (2009), 200–232. 10.1257/app.1.4.200

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh, "Inference under Covariate-Adaptive Randomization," *Journal of the American Statistical Association* 113:524 (2018), 1784–1796. 10.1080/01621459.2017.1375934

——— "Inference under Covariate-Adaptive Randomization with Multiple Treatments," *Quantitative Economics* 10:4 (2019), 1747–1785. 10.3982/QE1150

Christensen, Garret, and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56:3 (2018), 920–980. 10.1257/jel.20171350

Cochran, William G., and Gertrude M. Cox, *Experimental Designs* (New York: John Wiley & Sons, 1957).

Cohen, Jessica, and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics* 125:1 (2010), 1–45. 10.1162/qjec.2010.125.1.1

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Education, HIV, and Early Fertility: Experimental Evidence from Kenya," *American Economic Review* 105:9 (2015), 2757–2797. 10.1257/aer.20121607

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit" (pp. 3895–3962), in T. Paul Schultz and John A. Strauss (eds.), *Handbook of Development Economics*, Vol. 4 (Oxford: Elsevier, 2007). 10.1016/S1573-4471(07)04061-2

Elliott, Graham, Ulrich K. Müller, and Mark W. Watson, "Nearly Optimal Tests When a Nuisance Parameter Is Present under the Null Hypothesis," *Econometrica* 83:2 (2015a), 771–811. 10.3982/ECTA10535

——— "Supplement to 'Nearly Optimal Tests When a Nuisance Parameter Is Present under the Null Hypothesis'," *Econometrica* Supplemental Material (2015b).

Fisher, R. A., "The Arrangement of Field Experiments" (pp. 82–91), in Samuel Kotz and Norman L. Johnson (eds.), *Breakthroughs in Statistics: Methodology and Distribution* (New York: Springer, 1992).

Fisman, Raymond, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson, "Racial Preferences in Dating," *Review of Economic Studies* 75:1 (2008), 117–132. 10.1111/j.1467-937X.2007.00465.x

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science* 345:6203 (2014), 1502–1505. 10.1126/science.1255484

Gelman, Andrew, and John Carlin, "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science* 9:6 (2014), 641–651. 10.1177/1745691614551642

Gerber, A. S., and D. P. Green, *Field Experiments: Design, Analysis, and Interpretation* (New York: W. W. Norton, 2012).

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M. Lucas, and Derek Neal, "Educator Incentives and Educational Triage in Rural Primary Schools," *Journal of Human Resources* 57:1 (2022), 79–111. 10.3368/jhr.57.1.1118-9871R2

Gneezy, Uri, Kenneth L. Leonard, and John A. List, "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society," *Econometrica* 77:5 (2009), 1637–1664. 10.3982/ECTA6690

Hansen, Bruce E., *Econometrics* (Princeton, NJ: Princeton University Press, 2022).

Imbens, Guido W., and Charles F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica* 72:6 (2004), 1845–1857. 10.1111/j.1468-0262.2004.00555.x

Imbens, Guido W., and Donald B. Rubin, "Stratified Randomized Experiments" (pp. 187–218), in *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (New York: Cambridge University Press, 2015).

Kahan, Brennan C., "Bias in Randomised Factorial Trials," *Statistics in Medicine* 32:26 (2013), 4540–4549. 10.1002/sim.5869

Karlan, Dean, and John A. List, "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review* 97:5 (2007), 1774–1793. 10.1257/aer.97.5.1774

Karlan, Dean, and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," *Econometrica* 77:6 (2009), 1993–2008. 10.3982/ECTA5781

Kassler, Daniel, Ira Nichols-Barrer, and Mariel Finucane, "Beyond Treatment Versus Control: How Bayesian Analysis Makes Factorial Experiments Feasible in Education Research," *Evaluation Review* 44 (2019), 238–261. 10.1177/0193841X18818903

Kerwin, Jason T., and Rebecca L. Thornton, "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," this REVIEW 103:2 (2021), 251–264.

Ketz, Philipp, and Adam McCloskey, "Short and Simple Confidence Intervals When the Directions of Some Effects Are Known," this REVIEW (2023), 1–44.

Kohavi, Ron, Diane Tang, and Ya Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge: Cambridge University Press, 2020).

Kremer, Michael, "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons," *American Economic Review* 93:2 (2003), 102–106. 10.1257/000282803321946886

Leeb, Hannes, and Benedikt M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21:1 (2005), 21–59. 10.1017/S0266466605050036

——— "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *Annals of Statistics* 34 (2006), 2554–2591.

——— "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory* 24:02 (2008), 338–376.

List, John A., Sally Sadoff, and Mathis Wagner, "So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design," *Experimental Economics* 14:4 (2011), 439. 10.1007/s10683-011-9275-7

List, John A, Azeem M. Shaikh, and Yang Xu, "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics* 22 (2019), 773–793. 10.1007/s10683-018-09597-5

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania," *Quarterly Journal of Economics* 134:3 (2019), 1627–1673. 10.1093/qje/qjz010

McCloskey, Adam, "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," *Journal of Econometrics* 200 (2017), 17–35. 10.1016/j.jeconom.2017.05.001

——— "Asymptotically Uniform Tests after Consistent Model Selection in the Linear Regression Model," *Journal of Business & Economic Statistics* 38:4 (2020), 810–825.

Muralidharan, Karthik, and Venkatesh Sundararaman, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy* 119:1 (2011), 39–77. 10.1086/659655

Olken, Benjamin A., "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy* 115:2 (2007), 200–249. 10.1086/517935

Ray, D., *Development Economics* (Princeton, NJ: Princeton University Press, 1998).

Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66:5 (1974), 688. 10.1037/h0037350

Stoye, Jörg, "More on Confidence Intervals for Partially Identified Parameters," *Econometrica* 77:4 (2009), 1299–1315. 10.3982/ECTA7347

van der Vaart, A. W., *Asymptotic Statistics* (New York: Cambridge University Press, 1998).

Wasserstein, Ronald L., and Nicole A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *American Statistician* 70:2 (2016), 129–133. 10.1080/00031305.2016.1154108

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar, "Moving to a World beyond $p < 0.05$," *American Statistician* 73:Supp. 1 (2019), 1–19. 10.1080/00031305.2019.1583913

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics* 134:2 (2018), 557–598. 10.1093/qje/qjy029