Factorial designs, model selection, and (incorrect) inference in randomized experiments*

Karthik Muralidharan† Mauricio Romero‡ Kaspar Wüthrich§

July 18, 2019

Abstract

Cross-cutting or factorial designs are widely used in field experiments, with the aim of studying the effects of multiple treatments in a cost-effective way. However, this approach relies on the assumption that there are no interactions between treatments. We show that this assumption is problematic and leads to a significant increase in incorrect inference. Our re-analysis of data from all field experiments with factorial designs recently published in top-5 journals finds that 54% of treatment effects that are reported to be significant are no longer so after accounting for interactions. Further, pre-testing the interactions and ignoring them if they are not significant also leads to incorrect inference (due to the model selection implied by this procedure). While estimating the fully-saturated model (with all the interactions) corrects the problem, it comes at the cost of reduced power. We discuss three feasible alternatives to mitigate this trade-off: A Bonferroni-style correction, a nearly optimal test, and incorporating prior knowledge. We show that these corrections can yield greater power over a fully saturated model. But we also show that power to detect main effects will often be higher in designs that reallocate sample from the "interaction" cell to the main treatment cells. Thus, if the interactions are not of direct interest, it may be best to avoid cross-cutting designs altogether and use simpler non-interacted experimental designs.

Keywords: Randomized Controlled Trial; Factorial designs; Cross-cut designs; Field experiments
JEL Codes: C12, C18, C90, C93

*We are grateful to Arun Chandrasekhar, Clement de Chaisemartin, Stefano DellaVigna, Esther Duflo, Pascaline Dupas, Graham Elliott, Markus Goldstein, Hiroaki Kaido, Craig McIntosh, Adam McCloskey, Ben Olken, Gautam Rao, Andres Santos, and several seminar participants for comments and suggestions. We are also grateful to the authors of the papers we re-analyze for answering our questions and fact-checking that their papers are characterized correctly. 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.

†Department of Economics, UC San Diego; NBER; J-PAL; E-mail: kamurali@ucsd.edu
‡Centro de Investigación Económica, ITAM, Mexico City, Mexico; E-mail: mtromero@itam.mx
§Department of Economics, UC San Diego; E-mail: kwuthrich@ucsd.edu

1 Introduction

Cross-cutting or factorial designs are widely used in field experiments to study the effects of multiple treatments in a cost-effective way. However, unbiased estimation and correct inference of the main treatment effects in such experiments depend crucially on the assumption that the interaction between programs is negligible. 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".

This paper is motivated by the observation that many field experiments seem to be ignoring this caveat. 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 for their interaction) or a "short" model (only including dummies for both treatments). The long model yields consistent estimators for the average treatment effect of both treatments, as well as the interaction, and is always correct for inference regardless of the true value of the interaction. However, if the true value of the interaction effect is zero, the short model has greater power for conducting inference on the main treatment effects. This is why researchers often focus on presenting results from the short model - estimated without the interaction term, with the implicit assumption that the interaction is zero.

These gains in power, however, come at the cost of an increased likelihood of incorrect inference if the interaction effect is not zero.[1] We classify 27 out of 124 field experiments published in top-5 economics journals during 2006–2017 as using cross-cutting designs. Out of these 27 papers, 16 do not include all interaction terms in the main specifications. We re-analyzed the data from these papers by also including the interaction term.[2] Doing so has non-trivial implications for inference regarding the main treatment effects. The median change in the point estimates of the main treatment effects is 103%; about 26% of estimates change sign; 25% of the estimates switch significance at the 10% level, and 21% do so at the 5% level. Furthermore, 54% (26 out of 48) of estimates reported to be significant at the 5% level are no longer so after including interactions.[3]

[1] The incorrect inference is for testing treatment effects relative to a "business as usual" counterfactual.

[2] The full list of 27 papers ordered by citations is presented in Table A.1 in Appendix A. We re-analyzed 14 out of the 16 that do not include all interactions in the main specification. The other two papers did not have publicly-accessible data.

[3] The problem is more pronounced in papers with large numbers of treatments and interactions. Excluding the 3 papers with over 20 interaction terms, the rate of false rejection falls to 36% - out of 28 estimates reported to be significant at the 5% level, 10 are no longer so after including all interactions.

In practice, researchers often try to address the issue of interactions by conducting a two-step procedure where they first estimate the long model 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, the distribution of the estimators obtained from this data-dependent model selection procedure are complicated and highly non-normal, making the usual t-statistics misleading (Leeb & Pötscher, 2005, 2006, 2008). Further, cross-cutting experiments are *rarely adequately powered* to detect significant interactions.[4] Thus, the two-step procedure will almost always fail to reject that the interaction term is zero, even when it is different from zero. We show that, as a result, the two-step procedure will typically not control size, and thus often lead to incorrect inference regarding treatment effects against a "business as usual" counterfactual.

Textbook treatments of factorial designs Cochran & Cox (1957); Gerber & Green (2012) as well as Kremer (2003) are careful to clarify that the treatment effect in such designs (using the short model) should be interpreted as being conditional on the distribution of the other treatment arms in the experiment. Indeed, the issue that we highlight would not be a concern if we redefined the parameter of interest to be not the main treatment effect, but the composite treatment effect that includes a weighted-average of the interactions with other treatments. In practice, however, experimental results are rarely presented and interpreted with this important caveat.[5]

This status quo is problematic for at least three reasons. First, ignoring interactions affects the internal validity of experimental estimates. 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 in a study population depends on what other interventions are being studied in the *same* experiment, ignoring interactions is a threat to internal validity.[6] Policy discussions typically assume a "business as usual" counterfactual, and a treatment effect that is conditional on the nature and distribution of the other treatments in the same experiment will not usually represent a comparison with the policy-relevant counterfactual.

[4]For example, Gelman (2018) shows that you need 16 times the sample size to detect an interaction than to detect a main effect when the interactions are half the size of the main effects.

[5]Of the 16 papers in Table A.1 in Appendix A that present results from the short model without all interactions, we did not find any study that mentioned (in the main text or in a footnote) that the treatment effects should be interpreted this way (or against a counterfactual that was not "business as usual" but one that also had the other treatments in the same experiment).

[6]Further, researchers who use cross-cutting designs often use the two-step procedure described above, and present results from the short model *after* mentioning that the interactions are not significantly different from zero (see for example, Banerjee et al. (2007) and Karlan & List (2007)). This suggests that our view that interactions matter for internal validity is shared broadly.

Second, interactions are quantitatively important and typically not second-order. The median absolute magnitude of the interactions we estimate based on our re-analysis is 0.065σ and the median absolute value of the interactions relative to the main treatment effects is 0.37.[7] The view that interactions are second-order may have been influenced in part by the lack of evidence of significant interactions in most experiments to date. However, this is at least partly because very few experiments are adequately powered to detect interactions. Thus, "absence of evidence" of significant interactions may be getting erroneously interpreted as "evidence of absence." There is now both experimental (Mbiti et al., 2019; Duflo et al., 2015a) and non-experimental (Kerwin & Thornton, 2017; Gilligan et al., 2018) evidence that interactions matter. Indeed there is a long tradition in development economics that has highlighted the importance of complementarities across policies/programs in alleviating poverty traps (Ray, 1998; Banerjee & Duflo, 2005), which suggests that assuming away interactions in empirical work may be a mistake. As we show in Section 4, based on an example from (Mbiti et al., 2019), ignoring interactions would incorrectly lead to a conclusion that the main treatment effects were significant, when in fact, it is the interaction that drives the results.

Third, experimental evaluations typically inform policy through systematic reviews and meta-analyses of individual studies. These reviews often simply count the number of studies where an intervention has been found to be effective at conventional significance levels. Thus, the sensitivity of the significance of point estimates to the inclusion/exclusion of interaction terms and the extent of interaction effects with *other* treatments in the same study, is likely to have non-trivial implications for how evidence is summarized and translated into policy.

Given the issues with interpretation of estimates from the short model, we recommend that all cross-cutting experiments report the results from the long model. Further, the justification for the short model should not be that the interactions were not significant in the long model (because of the model selection issue discussed above). Rather, if researchers would like to present and discuss results from the short model, they should clearly indicate that the presented treatment effects should be interpreted as being conditional on the distribution of other treatments in the same experiment. This will ensure transparency in interpretation of the main results as including interactions, and enable readers to assess the extent to which the other treatments may or may not be typical background factors which can be ignored.

[7] The median value of the interaction term across studies is in fact close to zero. The problem is that the median *absolute* value of the interaction term is not zero, which results in a non-trivial rate of false rejection of the null hypothesis *in any given study*.

The discussions in Kremer (2003) and Duflo et al. (2007) suggest that an important motivation for cross-cutting designs is the belief that interactions are "small" relative to the main treatment effects of interest. We therefore consider if it may be possible to design tests for the main treatment effects relative to a "business as usual" counterfactual that improve power relative to the t-test based on the long model *while maintaining size control*. Based on recent advances in theoretical econometrics, we identify and explore three promising possibilities.

The first is to control size in the presence of model selection using a Bonferroni approach. This amounts to constructing an asymptotically valid confidence interval for the interaction, and then search for the maximal critical value of the t-test in the confidence interval. The uncertainty around the test statistic can be accounted for using a Bonferroni-correction (McCloskey, 2017, 2019). Overall, our simulations suggest that while this procedure adequately controls for size, it can be conservative and may not improve power relative to the long model.

The second approach targets power towards a likely value of the interaction (e.g., near an interaction of zero), while controlling size for all values of the interaction. The idea is to view the inference problem as a testing problem in the presence of a nuisance parameter (the interaction effect) and to adopt the standard approach of integrating the parameter of interest with respect to some probability distribution under the null and under the alternative. The resulting test is then simply a likelihood ratio test and the task becomes to find the probability distribution under the null which yields the optimal test, the so-called least favorable distribution. Because the least favorable distribution is challenging to identify, Elliott et al. (2015) propose to focus on an approximate least favorable distribution, which is determined numerically and yields a nearly optimal test. Our simulation evidence suggests that the nearly optimal test achieves local power improvements relative to the long model near the likely value of the interaction, but can have much lower power further away from this value.

A third approach is to construct valid confidence intervals for the main treatment effects under prior knowledge of the interaction. Specifically, researchers may assume that the interaction is no larger than a pre-specified fraction of the minimum main effect that the experiment is powered to detect. Under this assumption, the main effects are no longer point but only partially or set identified and we construct confidence intervals based on Imbens & Manski (2004) and Stoye (2009). We confirm in simulations that, when the prior knowledge is correct, this approach controls size, and also yields power gains relative to the long model. However, it may suffer from size distortions if the prior knowledge is incorrect. Since the problem we identify is mainly a result of not knowing

the value of the interaction *ex ante*, this approach may be of limited use in practice.

Our recommendation for the analysis of completed experiments with 2×2 factorial designs is to use the tests based on Elliott et al. (2015), and we have written the code needed to implement this procedure to facilitate adoption (`https://mtromero.shinyapps.io/elliott/`). However, this procedure is computationally prohibitive for cross-cutting designs with multiple interaction cells, because these designs introduce multiple nuisance parameters (interaction terms) into the inference problem. Thus it cannot be recommended for complicated factorial designs, where the only viable option may be to report the results from the long model, or to suitably caveat the interpretation of results from the short model as discussed above.

All three approaches above control size in *existing* cross-cutting experiments. But for the design of new experiments, a natural alternative is to leave the "interaction cell" empty and increase the number of units assigned exclusively to one of the treatments or the control group. Our simulations show that leaving the interaction cell empty yields more power gains than any of the econometric methods discussed above for most of the relevant values of the interaction effect.[8] Thus, if one is not interested in the interaction between the programs, we suggest avoiding factorial designs. If interactions are of research interest, the experiment should be powered to detect them.

Our first contribution is to the literature on the design of field experiments. Athey & Imbens (2017), Bruhn & McKenzie (2009), and List et al. (2011) provide guidance on the design of field experiments, but do not discuss when and when not to implement factorial designs. Duflo et al. (2007) 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 re-analysis of existing experiments as well as simulations suggests that *there is no free lunch* and that the perceived gains in power and cost-effectiveness from running experiments with 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 including sets of interactions with other treatments that may not exist in a typical counterfactual scenario.

We also contribute to the literature that aims at improving the econometric analysis of completed field experiments. Two notable recent examples are Young (in press), who shows that randomization tests of the significance result in 13% to 22% fewer significant

[8]This design will yield power gains relative to running two separate experiments, because the control group is used twice. But it avoids the problem of interactions noted above. An example of such a design is provided by Muralidharan & Sundararaman (2011) who study the impact of four different interventions in one experiment with one common control group, but no cross-cutting treatment arms.

results than those originally reported in the paper, and List et al. (2016) who present a procedure to correct for multiple hypothesis testing in field experiments. 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 be brought to bear to improve inference in factorial designs which are ubiquitous in economics field experiments.

The rest of this paper is organized as follows. Section 2 describes selected theoretical aspects of cross-cut designs. Section 3 examines the use of these designs in papers published in the top-5 economics journals, and presents our re-analysis of these studies after including the interaction term. Section 4 shows that ignoring the interaction and model selection (pre-testing) leads to incorrect inference and false positives. Section 5 examines three econometric approaches to improve power in factorial designs while controlling size. Section 6 discusses whether factorial designs make sense for new studies, Section 7 considers factorial designs with more than two treatments, and Section 8 concludes.

2 Theoretical aspects of cross-cut designs

In this section, we discuss identification, estimation and inference in experiments with factorial designs. For simplicity, we focus on factorial designs with two treatments, T_1 and T_2 (commonly known as 2×2 designs), where a researcher randomly assigns some subjects to receive treatment T_1, some subjects to receive treatment T_2, and some subjects to receive both treatments (see Table 1). It is straightforward to extend the analysis to cross-cut designs with more than two treatments; see Section 7.

Table 1: 2×2 factorial design

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

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

2.1 Setup and effects of interest

We formalize the problem using the potential outcomes framework of Rubin (1974). We are interested in identifying and estimating the causal effect of the two treatments, T_1

and T_2, on an outcome of interest, Y. In our setting, 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

$$Y = Y_{0,0} \cdot \mathbf{1}_{\{T_1=0,T_2=0\}} + Y_{1,0} \cdot \mathbf{1}_{\{T_1=1,T_2=0\}} + Y_{0,1} \cdot \mathbf{1}_{\{T_1=0,T_2=1\}} + Y_{1,1} \cdot \mathbf{1}_{\{T_1=1,T_2=1\}}, \quad (1)$$

where $\mathbf{1}_{\{A\}}$ is an indicator function which is equal to one if the event A is true and zero otherwise. There are different types of average treatment effects (ATE):

$E(Y_{1,0} - Y_{0,0})$: ATE of T_1 relative to a counterfactual where $T_2 = 0$
$E(Y_{0,1} - Y_{0,0})$: ATE of T_2 relative to a counterfactual where $T_1 = 0$
$E(Y_{1,1} - Y_{0,1})$: ATE of T_1 relative to a counterfactual where $T_2 = 1$
$E(Y_{1,1} - Y_{1,0})$: ATE of T_2 relative to a counterfactual where $T_1 = 1$
$E(Y_{1,1} - Y_{0,0})$: ATE of T_1 and T_2 combined relative to a counterfactual where $T_1 = T_2 = 0$

In what follows, we will refer to $E(Y_{1,0} - Y_{0,0})$ and $E(Y_{0,1} - Y_{0,0})$ as the *main effects* (relative to a "business as usual" counterfactual) of T_1 and T_2. The interaction effect, i.e., the difference between the joint effect of combining both treatments and the sum of the main effects, is

$$\begin{aligned} E(Y_{1,1} - Y_{0,0}) - [E(Y_{0,1} - Y_{0,0}) + E(Y_{1,0} - Y_{0,0})] &= E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}) \\ &= E(Y_{1,1} - Y_{0,1}) - E(Y_{1,0} - Y_{0,0}) \\ &= E(Y_{1,1} - Y_{1,0}) - E(Y_{0,1} - Y_{0,0}) \quad (2) \end{aligned}$$

Equation (2) shows that the interaction effect is equal to the difference between the ATE of T_1 relative to the counterfactual where $T_2 = 1$ and the ATE of T_2 relative to a counterfactual where $T_1 = 0$. Alternatively, the interaction effect can be expressed as the difference between the ATEs of T_2 relative to counterfactuals where $T_1 = 1$ and $T_1 = 0$.

We assume that both treatments are randomly assigned and independent of each other.

Assumption 1. $(Y_{0,0}, Y_{1,0}, Y_{0,1}, Y_{1,1}) \perp\!\!\!\perp (T_1, T_2)$ *as well as* $(Y_{0,0}, Y_{1,0}, Y_{0,1}, Y_{1,1}, T_1) \perp\!\!\!\perp T_2$ *and* $(Y_{0,0}, Y_{1,0}, Y_{0,1}, Y_{1,1}, T_2) \perp\!\!\!\perp T_1$

Assumption 1 and the definition of potential outcomes (see Equation (1)) imply that, for $(t_1, t_2) \in \{0,1\} \times \{0,1\}$,

$$\begin{aligned} E(Y \mid T_1 = t_1, T_2 = t_2) &= E(Y_{t_1,t_2} \mid T_1 = t_1, T_2 = t_2) \\ &= E(Y_{t_1,t_2}). \end{aligned}$$

As a consequence, the different ATEs are identified as

$$\begin{aligned}
E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0) &= E(Y_{1,0} - Y_{0,0}) \\
E(Y \mid T_1 = 0, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 0) &= E(Y_{0,1} - Y_{0,0}) \\
E(Y \mid T_1 = 1, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 1) &= E(Y_{1,1} - Y_{0,1}) \\
E(Y \mid T_1 = 1, T_2 = 1) - E(Y \mid T_1 = 1, T_2 = 0) &= E(Y_{1,1} - Y_{1,0}) \\
E(Y \mid T_1 = 1, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 0) &= E(Y_{1,1} - Y_{0,0})
\end{aligned}$$

and the interaction effect is identified via Equation (2).

2.2 Short and long regression models

In Section 3 we document that researchers analyzing cross-cut designs typically consider one of the following two population regression models:

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon, \quad \text{(long model)} \tag{3}$$

$$Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s \quad \text{(short model)} \tag{4}$$

The "long" or fully saturated model (3) includes both treatment indicators as well as their interaction. By contrast, the "short" model (4) only includes the two treatment indicators but ignores the interaction term.

Next, we relate the population regression coefficients in these models to the causal effects defined in Section 2.1.[9] For the fully saturated model (3), we obtain

$$\begin{aligned}
\beta_1 &= E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0), \\
\beta_2 &= E(Y \mid T_1 = 0, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 0), \\
\beta_{12} &= E(Y \mid T_1 = 1, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 1) \\
&\quad - [E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0)].
\end{aligned}$$

Under Assumption 1, the results in Section 2.1 imply that

$$\beta_1 = E(Y_{1,0} - Y_{0,0}), \tag{5}$$

$$\beta_2 = E(Y_{0,1} - Y_{0,0}), \tag{6}$$

$$\beta_{12} = E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}). \tag{7}$$

[9] The population regression coefficient β in the model $Y = X'\beta + \varepsilon$ is the solution to the population least squares problem and is given by $\beta = E(XX')^{-1} E(XY)$.

Equations (5)–(7) show that the population regression coefficients in the long regression model correspond to the main effects of T_1 and T_2 and the interaction effect.

Under Assumption 1, the population regression coefficients in the short model (4) are

$$\begin{aligned} \beta_1^s &= E(Y \mid T_1 = 1) - E(Y \mid T_1 = 0) && (8) \\ \beta_2^s &= E(Y \mid T_2 = 1) - E(Y \mid T_2 = 0) && (9) \end{aligned}$$

One can derive the following two expressions for β_1^s and β_2^s (see Appendix C.1 for details):

$$\begin{aligned} \beta_1^s &= \beta_1 + \beta_{12} P(T_2 = 1) \\ &= E(Y_{1,0} - Y_{0,0}) + E(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0})\, P(T_2 = 1) && (10) \\ \beta_2^s &= \beta_2 + \beta_{12} P(T_1 = 1) \\ &= E(Y_{0,1} - Y_{0,0}) + E(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0})\, P(T_1 = 1) && (11) \end{aligned}$$

and

$$\begin{aligned} \beta_1^s &= (\beta_{12} + \beta_1)\, P(T_2 = 1) + \beta_1 P(T_2 = 0) \\ &= E(Y_{1,1} - Y_{0,1})\, P(T_2 = 1) + E(Y_{1,0} - Y_{0,0})\, P(T_2 = 0) && (12) \\ \beta_2^s &= (\beta_{12} + \beta_2)\, P(T_1 = 1) + \beta_2 P(T_1 = 0) \\ &= E(Y_{1,1} - Y_{1,0})\, P(T_1 = 1) + E(Y_{0,1} - Y_{0,0})\, P(T_1 = 0) && (13) \end{aligned}$$

In Equation (10), β_1^s is equal to the sum of the ATE of T_1 relative to a counterfactual where $T_2 = 0$ and the interaction effect multiplied by the fraction of individuals with $T_2 = 1$. Alternatively, β_1^s can be written as a convex combination of the ATE of T_1 relative to a counterfactual where $T_2 = 1$ and the ATE of T_1 relative to a counterfactual where $T_2 = 0$; see Equation (12). The weights correspond to the fractions of individuals with $T_2 = 1$ and $T_2 = 0$. Equations (11) and (13) present the corresponding expressions for β_2^s.

This analysis shows that, unless the interaction effect is truly zero (in which case $\beta_1 = \beta_1^s$ and $\beta_2 = \beta_2^s$), the population regression coefficients in the short regression model differ from the corresponding coefficients in the long model and neither correspond to the main effects nor the interaction effect.

2.3 Estimation and inference

Suppose that the researcher has access to a random sample $\{Y_i, T_{1i}, T_{2i}\}_{i=1}^N$. Consider a factorial design with sample sizes as in Table 1. Under Assumption 1 and standard regularity conditions, the OLS estimators based on the long regression model are consistent for the main effects and the interaction effect: $\hat{\beta}_1 \xrightarrow{p} \beta_1 = E(Y_{1,0} - Y_{0,0})$, $\hat{\beta}_2 \xrightarrow{p} \beta_2 = E(Y_{0,1} - Y_{0,0})$ and $\hat{\beta}_{12} \xrightarrow{p} \beta_{12} = E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0})$. By contrast, the probability limits of the OLS estimators based on the short model are given by

$$\begin{aligned} \hat{\beta}_1^s &\xrightarrow{p} \beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1), \\ \hat{\beta}_2^s &\xrightarrow{p} \beta_2^s = \beta_1 + \beta_{12} P(T_2 = 1). \end{aligned}$$

Unless the true interaction effect is zero (i.e., $\beta_{12} = 0$), $\hat{\beta}_1^s$ and $\hat{\beta}_2^s$ are not consistent estimators of the main effects. Thus, if the goal is consistency, one should always use the long model.

The choice between the long and the short regression model is not as clear cut when it comes to inference. To illustrate, suppose that the data generating process is given by

$$Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} \times T_{2i} + \varepsilon_i,$$

where $\varepsilon_i \mid T_{1i}, T_{2i} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Normality allows us to formally compute and compare the finite sample power of the t-tests based on the short and the long regression model. In what follows, we focus on β_1. The analysis for β_2 is symmetric and thus omitted.

If the interaction effect is zero (i.e., $\beta_{12} = 0$), it follows from standard results that, conditional on $(T_{11}, \ldots, T_{1N}, T_{21}, \ldots, T_{2N})$, $\hat{\beta}_1 \sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right)$ and $\hat{\beta}_1^s \sim N\left(\beta_1, Var\left(\hat{\beta}_1^s\right)\right)$, where

$$Var\left(\hat{\beta}_1\right) \equiv \sigma^2 \frac{N_1 + N_2}{N_1 N_2} \quad \text{and} \quad Var\left(\hat{\beta}_1^s\right) \equiv \sigma^2 \frac{N_1 N_3 + N_1 N_4 + N_2 N_3 + N_2 N_4}{N_1 N_2 N_3 + N_1 N_2 N_4 + N_1 N_3 N_4 + N_2 N_3 N_4}.$$

Because the variance of $\hat{\beta}_1$ is larger than the variance of $\hat{\beta}_1^s$, the t-test based on the short model exhibits higher finite sample power than the t-test based on the long model.[10] Appendix D contains a formal proof of this claim as well as a brief discussion of how the power gains vary as a function of the "size" of the interaction cell, N_4.

[10] To see this, note that

$$Var\left(\hat{\beta}_1\right) - Var\left(\hat{\beta}_1^s\right) = \sigma^2 \frac{N_3 N_4 (N_1 + N_2)^2}{N_1 N_2 (N_1 N_2 N_3 + N_1 N_2 N_4 + N_1 N_3 N_4 + N_2 N_3 N_4)} \geq 0.$$

If, on the other hand, the interaction effect is not zero (i.e., $\beta_{12} \neq 0$), ignoring the interaction can lead to substantial size distortions as we demonstrate in Section 4.2. Depending on the true value of the interaction effect, the finite sample power of the t-tests based on the short model can be higher or lower than the power of the t-tests based on the long model.

3 Factorial designs in practice

In this section we document common practices among researchers studying field experiments with factorial designs. We analyze all articles published between 2006 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 E.1 provides more details). The *Quarterly Journal of Economics* (QJE) has both the highest absolute (45) and relative number of field experiments (9% of articles published in the QJE in this period). Factorial designs are widely used: Among field experiments 27 (22%) had a factorial design.[12] Only 11 articles with factorial designs (∼41%) included all interaction terms in their main specification (see Table 2).

Table 2: Field experiments published in top-5 journals between 2006 and 2017

	Statistic
Field experiments published	124
Field experiments with factorial designs	27
All interactions included in the main specification	11
Average number of treatments per paper	1.6
Average number of treatments per paper | factorial design	2.9

[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" which consist of papers presented at the annual meeting of the American Economics Association.

[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 re-randomized to study behavioral changes conditional on a realization of the previous treatment. Examples of studies with two-stage randomization designs include Cohen & Dupas (2010), Karlan & Zinman (2009), and Ashraf et al. (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), Gneezy et al. (2009), and Fisman et al. (2008)).

3.1 Effect of ignoring interactions

What is the effect of omitting interaction terms in the main specification? To answer this question, we re-analyze the data from all field experiments with factorial designs and publicly available data that do not include all the interactions in the main specification. Of the ten most-cited papers with factorial designs listed in Table A.1 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-5 journals, 16 papers do not include all interaction terms. Appendix A provides details of the experimental design of each of the 16 papers. Of these, 2 papers did not have publicly-available replication data.

We downloaded the publicly-available data files and replicated the main results in each of the remaining 14 papers. We then compared the original treatment effects (estimated without the interaction terms) with those estimated including the interaction terms. In other words, we compare estimates based on Equation (4) versus those based on Equation (3)).

Figure 1 presents summary statistics on how the original treatment estimates compare to the estimates after accounting for the interaction terms. The median absolute change in the point estimate of the main treatment effect is 103% when it is estimated using the long model compared to the short model. Roughly 26% of estimated treatment effects change sign when they are estimated using the long regression.

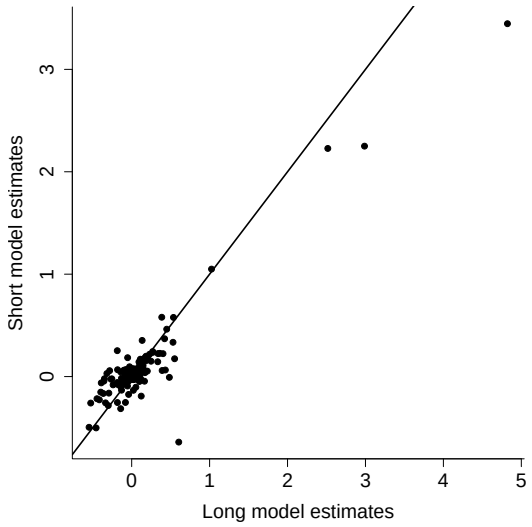
Table 3 shows how the significance of the main treatment estimates change when using the long and the short model. About 49% of treatment estimates that were significant at the 10% level when using the short model are no longer significant when the long model is estimated. 54% and 56% of estimates lose 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 when the long regression is estimated (7%, 6%, and 2%, at the 10%, 5%, and 1% levels respectively).

Similar results are seen when we restrict our re-analysis to the ten most cited papers with factorial designs that do not include the interaction terms. Corresponding figures and tables are presented in Appendix A.17.2 (Figure A.9 and Table A.2).

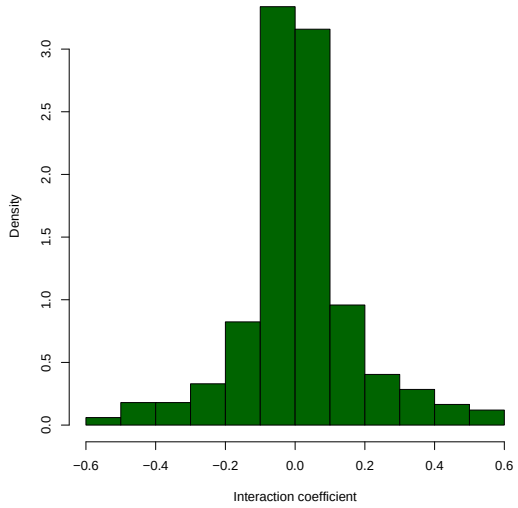
Finally, the problem of incorrect inference from ignoring interaction terms is more pronounced in papers with a large number of treatments and interactions. When we exclude the three papers with over 20 interactions in them, the problem is still substantial but of lower magnitude. Corresponding figures and tables are presented in Appendix A.17.3 (Figure A.11 and Table A.3).

Figure 1: Treatment estimates from the long and the short regression

(a) Main treatment estimates



(b) Interaction



Note: Both figures show treatment estimates from all papers with factorial designs and publicly available data that do not include the interaction in the original study. Figure 1a shows how the main treatment estimates change across the short and the long model across all studies. The median main treatment estimate from the short model is 0.013σ, the median main treatment estimate from the long model is 0.014σ, the average absolute difference between the treatment estimates of the short and the long model is 0.062σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 103%, and 26% of treatment estimates change sign when they are estimated using the long or the short model. Figure 1b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is 0.001σ, 6.3% of interactions are significant at the 10% level, 3.7% are significant at the 5% level, and 0.9% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.37.

Table 3: Significance of treatment estimates from the long and the short regression

Panel A: Significance at the 10% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	82	31	113
Significant	6	32	38
Total	88	63	151

Panel B: Significance at the 5% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	97	26	123
Significant	6	22	28
Total	103	48	151

Panel C: Significance at the 1% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	124	14	138
Significant	2	11	13
Total	126	25	151

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

4 Issues with ignoring the interaction and model selection

In Section 3 we document that many studies (including nearly all of the most highly-cited studies) do not include all the interaction terms. One possibility is that authors are using a two-step procedure to determine whether to focus on the full model or the short model, and only showing the full model if the interactions are significant. Indeed, only 6.3% of interactions are significant at the 10% level, 3.7% are significant at the 5% level, and 0.9% are significant at the 1% level among the 14 papers that we re-analyzed. Thus, the lack of inclusion of interactions may reflect authors' beliefs that the interactions are second order as seen by their lack of significance in the long model.

However, as we show, both of these approaches – ignoring the interaction, and model selection (or pre-testing) – are problematic for making inference about the main effects. Towards this end, we assess the finite sample performance of both approaches based on an application-based running example. For concreteness, we focus on the problem of testing the null hypotheses that β_1 is equal to zero, $H_0 : \beta_1 = 0$. The analysis for β_2 is symmetric and thus omitted.

4.1 Running example

To illustrate our results, we use a running example based on Mbiti et al. (2019) who study the effectiveness of alleviating (monetary) resource constraints in public primary schools, the effectiveness of introducing teacher performance pay, and the complementarities between the two. We use this experiment for the running example since it is explicitly powered to estimate the interaction effect. The experiment has a cross-cut design with two treatments and four groups. The first group gets only one treatment (inputs), the second group gets only the other treatment (incentives), the third group gets both treatments (inputs and incentives), and the fourth group serves as the control.

Based on Mbiti et al. (2019), our main results use a sample size of 1,100 (from an i.i.d. draw), evenly split across groups.[13] We use Monte Carlo simulations to assess the bias, size and power of different inference approaches. Specifically, we consider the following data generating process to create N observations:

$$Y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} \times T_{2i} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0,1)$ and T_{1i} and T_{2i} are randomly assigned treatments with $P(T_{1i} = 1) = P(T_{2i} = 1) = 0.5$. We perform simulations with 10,000 repetitions and report the bias and the rejection rates under the null (size) and the alternative hypothesis (power).

4.2 Ignoring the interaction

A first popular approach is to ignore the interaction and to focus on the short regression model. This procedure is justified if the researcher is certain that $\beta_{12} = 0$, in which case

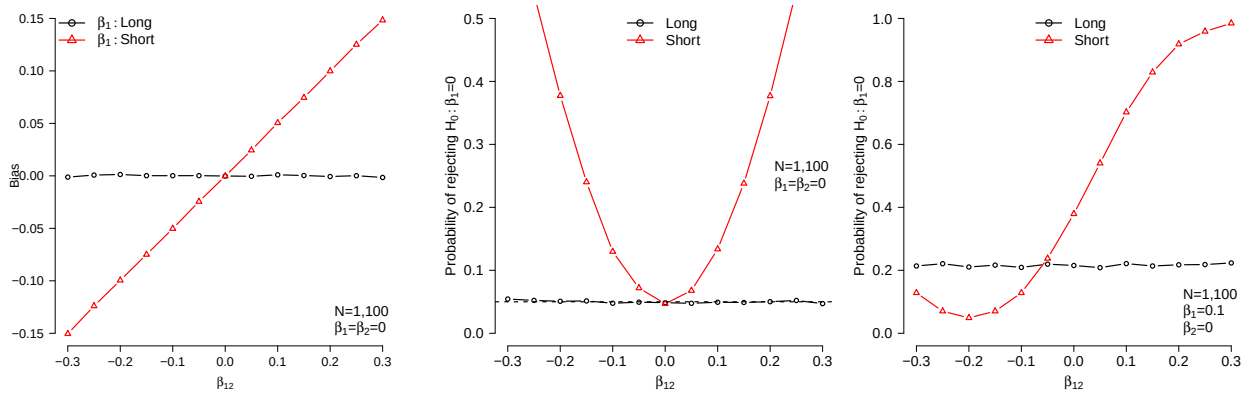
[13] The sample in Mbiti et al. (2019) was a representative set of 350 schools across 10 districts in Mainland Tanzania. 70 schools received Capitation Grants (per pupil grants), 70 schools received a simple performance pay program, 70 schools received both grants and incentives, and 140 schools served as the control group. 30 students were sampled from each school to create a panel of 10,500 students. Had the authors conducted an individual level-experiment, the sample size to achieve the same power would have been 1,100 (see Appendix B).

it leads to consistent estimates of β_1 and to power improvements relative to the long model (see Section 2.3). However, if $\beta_{12} \neq 0$, ignoring the interaction yields inconsistent estimates and size distortions.

To illustrate this trade-off, Figure 2 shows how bias, size and power vary across different values of β_{12} in both the long and the short model. As expected, the long model exhibits no bias and correct size for all values of β_{12}, while the short model has a bias and does not achieve size control when $\beta_{12} \neq 0$. The trade-off is that for $\beta_{12} = 0$, the short model controls size and exhibits higher finite sample power than the long model. When $\beta_{12} \neq 0$, the power of the t-test based on the short model depends on β_{12} and may be higher or lower than the power of tests based on the long model. The main takeaway from Figure 2 is that researchers should avoid the short model, unless there is no uncertainty that $\beta_{12} = 0$. In Mbiti et al. (2019), running the short model would have led to the incorrect conclusion that the teacher incentive program they studied was effective in all subjects in both years and that the school grant program was effective in the second year (see Table 4: Panel B).

Figure 2: Bias, size control and power trade-off

(a) Bias (b) Size (c) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 2b and 2c is $\alpha = 0.05$.

4.3 Model selection (or pre-testing)

In view of the power gains described in Section 2.3, a common practice is to employ a data-driven two-step procedure to determine whether to estimate the full model or to ignore the interaction. Specifically, the steps are:

1. Estimate the long model and test the null hypothesis that β_{12} is zero (i.e., $H_0 : \beta_{12} =$

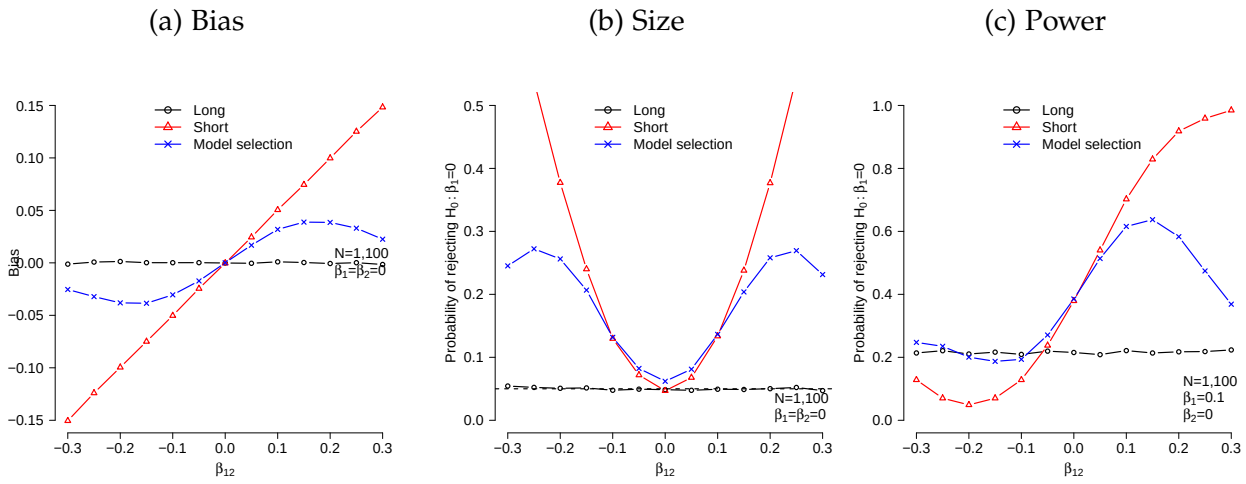
0) using a t-test.

2. (a) If $H_0 : \beta_{12} = 0$ is rejected, test $H_0 : \beta_1 = 0$ using the t-test based on the long model.

   (b) If $H_0 : \beta_{12} = 0$ is not rejected, test $H_0 : \beta_1 = 0$ using the t-test based on the short model.

It is well-known that the distributions of the estimators obtained from this data-dependent model selection procedure are complicated and highly non-normal, rendering the usual t-statistic-based inference invalid (e.g., Leeb & Pötscher, 2005, 2006, 2008). We illustrate this issue in the context of our running example. The bias, size and power using model selection are shown in Figure 3. For reference, we also include the long and the short model. The problem is particularly pronounced for the range of values of β_{12} that the experiment is not powered to detect.

Figure 3: Model selection is biased and does not control for size



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 3b and 3c 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.

The main takeaway from Figure 3 is that pre-testing leads to incorrect inference, and researchers should avoid it. In Mbiti et al. (2019), model selection would have led to the incorrect conclusion that teacher incentives were effective in the first year (see Table 4: Panel C). In this study, the design was powered to detect interactions greater than 0.1σ. Thus, the model selection would have correctly led the authors to use the long model in the second year.

However, in practice, most field experiments are not powered to detect meaningful interactions (especially in studies with multiple cross-cutting treatments). As illustrated

by (Gelman, 2018), a study would need 16 times the sample size to detect an interaction than to detect a main effect when the interactions are half the size of main effects. As a result, model selection will typically lead to incorrectly estimating the short model, and correspondingly lead to incorrect inference about the main effects.

Table 4: Treatment effects on test scores from Mbiti et al. (2019)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Year 1				Year 2			
	Math	Kiswahili	English	Combined (PCA)	Math	Kiswahili	English	Combined (PCA)
Panel A: Long model								
Incentives	0.06	0.05	0.06	0.06*	0.07*	0.01	0.00	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
Grants	-0.05	-0.01	-0.02	-0.03	0.01	-0.00	0.02	0.01
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
Incentives × Grants	0.10*	0.06	0.07	0.09	0.12*	0.20***	0.16*	0.18***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.08)	(0.07)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
Panel B: Short model								
Incentives	0.10***	0.08***	0.09***	0.10***	0.13***	0.10***	0.07*	0.11***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)
Grants	-0.01	0.02	0.01	0.01	0.06	0.08**	0.09**	0.09**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439
Panel C: Model selection								
Incentives	0.06	0.08***	0.09***	0.10***	0.07*	0.01	0.00	0.03
	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)
Grants	-0.05	0.02	0.01	0.01	0.01	-0.00	0.02	0.01
	(0.04)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
Incentives × Grants	0.10*				0.12*	0.20***	0.16*	0.18***
	(0.06)				(0.07)	(0.07)	(0.08)	(0.07)
N. of obs.	9,142	9,142	9,142	9,142	9,439	9,439	9,439	9,439

Results from estimating the long (Panel A) and the short (Panel B) model for the experiment conducted by Mbiti et al. (2019). Panel A estimates the long model. Panel B estimates the short model (ignoring the interaction). Panel C performs model selection, where the interaction is only included in the final estimation if its significant at the 10% level. Clustered standard errors, by school, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Can we improve power while achieving size control?

We now examine whether it is possible to improve power relative to the the t-test based on the long model, while maintaining size control across relevant values of β_{12}. We explore three different approaches: model selection with a Bonferroni correction, a nearly optimal test, and incorporating prior knowledge. We focus on 2×2 factorial designs; see Section 7 for a discussion of factorial designs with more than two treatments.

5.1 Model selection with Bonferroni corrections and nearly optimal tests

Here, we focus on β_1 and partial out T_2. To simplify the exposition, we define $T_{12} := T_1 \times T_2$ and keep the partialling-out implicit. The regression model of interest is[14]

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

We are interested in the follow hypothesis testing problem:

$$H_0: \ \beta_1 = 0, \ \ \beta_{12} \in \mathbb{R} \quad \text{against} \quad H_0: \ \beta_1 \neq 0, \ \ \beta_{12} \in \mathbb{R} \tag{14}$$

Both the null and the alternative hypothesis are composite so that the null specifies a single value for β_1 but not for β_{12}. In other words, the nuisance parameter (β_{12}) is present under the null hypothesis. The t-test based on the long regression model is known to be the uniformly most powerful test among tests that are unbiased for all $\beta_{12} \in \mathbb{R}$ (e.g., van der Vaart, 1998; Elliott et al., 2015).[15] Hence, if one restricts attention to this class of tests, then the optimal inference procedure is simply the standard t-test based on the long model.

Here we investigate the finite sample performance of two alternative inference approaches that may outperform the t-test based on the long model locally (i.e., have higher power in certain regions), while controlling size for all values of β_{12}:

1. Model selection combined with a Bonferroni-style correction (McCloskey, 2017)
2. A nearly optimal test targeting power towards an a priori likely value $\beta_{12} = \bar{\beta}_{12}$ (Elliott et al., 2015)

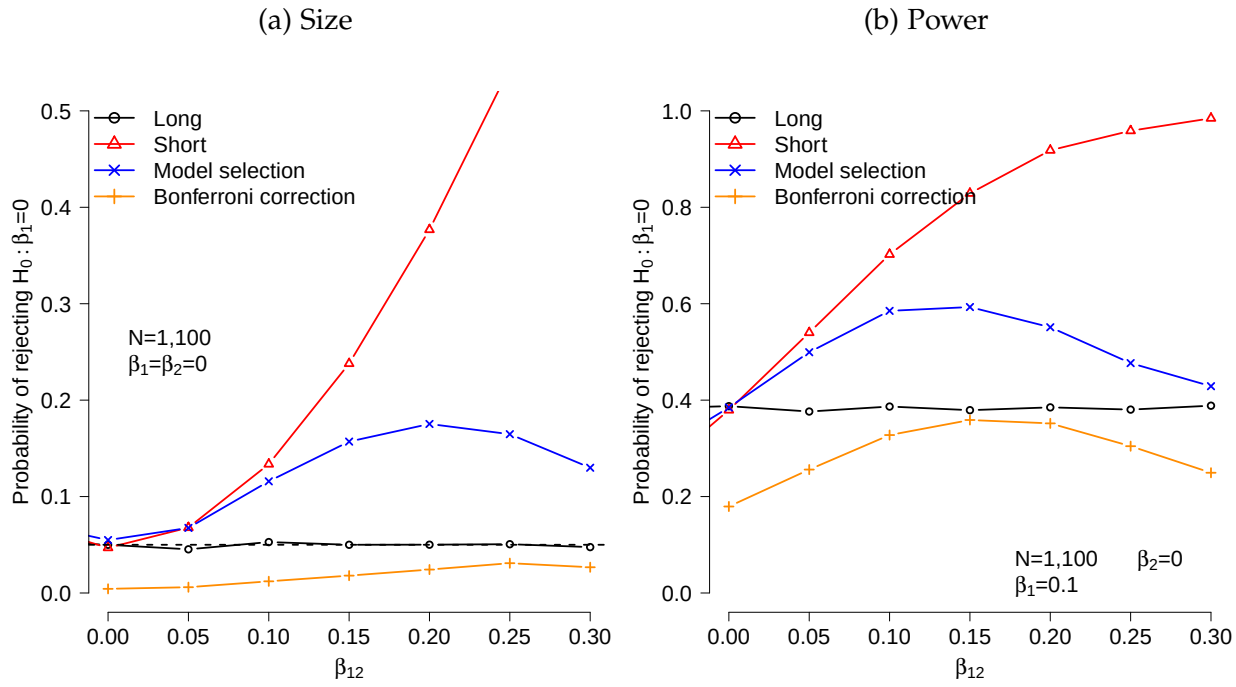
[14] We omit the intercept because, after partialling out T_2, all random variables have mean zero as residuals from a population regression.

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

5.1.1 Model Selection with a Bonferroni-style Correction

A natural approach to control size in the presence of model selection is to take a least favorable (LF) approach. That is to use the largest critical value across all values for the nuisance parameter (e.g., Andrews & Guggenberger, 2009; Leeb & Pötscher, 2017). However, it is well-known that this worst case approach can exhibit poor power properties. McCloskey (2017) suggests a procedure that improves upon the LF approach, asymptotically controls size and has non-negligible power. The basic insight of this approach is that one can construct an asymptotically valid confidence interval for β_{12}. As a consequence, one can search for the largest critical value over the values of β_{12} in the confidence interval rather than over the whole parameter space as in the LF approach. The uncertainty about the nuisance parameter (β_{12}) and the test statistic can be accounted for using a Bonferroni-correction. Alternatively, one can adjust critical values according to the null limiting distributions that arise under drifting parameter sequences. We refer to McCloskey (2017, 2019) for more details as well as specific implementation details.[16]

Figure 4: McCloskey (2017)'s Bonferroni-style correction controls size but does not exhibit power gains relative to the long model



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 4a and 4b 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.

[16] We implement the adjusted Bonferroni critical values outlined in Section 3.2 and use the algorithm "Algorithm Bonf-Adj" in the Appendix of McCloskey (2017). We employ conservative model selection and the use a tuning parameter of 0.9α, where α is the nominal level of the test.

Figure 4 reports the results of applying McCloskey (2017)'s Bonferroni-style correction to our running example. It shows that model selection with state-of-the-art Bonferroni adjustments leads to tests that control size for all values of β_{12}. However, this method can be conservative and may not lead to power gains relative to the t-test based on the long model, at least not over the regions of the parameter space considered here.[17]

5.1.2 Nearly Optimal Tests Targeting Power towards a likely Value $\beta_{12} = \bar{\beta}_{12}$

Consider a scenario where a particular value $\beta_{12} = \bar{\beta}_{12}$ seems a priori likely and suppose that we want to find a test that controls size 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 following testing problem

$$H_0 : \ \beta_1 = 0, \ \ \beta_{12} \in \mathbb{R} \quad \text{against} \quad H_1 : \ \beta_1 \neq 0, \ \ \beta_{12} = 0. \tag{15}$$

We use the numerical algorithm developed by Elliott et al. (2015, EMW henceforth) to construct a nearly optimal test for the testing problem (15). To describe their procedure, note that under standard conditions, the t-statistics are approximately normally distributed in large samples

$$\begin{pmatrix} \hat{t}_1 \\ \hat{t}_{12} \end{pmatrix} \sim N\left(\begin{pmatrix} t_1 \\ t_{12} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{16}$$

where $\hat{t}_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$, $\hat{t}_{12} = \frac{\hat{\beta}_{12}}{SE(\hat{\beta}_{12})}$, $t_1 = \frac{\beta_1}{SE(\hat{\beta}_1)}$, $t_{12} = \frac{\beta_{12}}{SE(\hat{\beta}_{12})}$, and $\rho = Cov(t_1, t_{12})$. We also define $\hat{t} = (\hat{t}_1, \hat{t}_{12})$ and $t = (t_1, t_{12})$. $SE(\hat{\beta}_1)$, $SE(\hat{\beta}_{12})$ and $Cov(t_1, t_{12})$ can be consistently estimated under weak conditions (here we use a standard heteroscedasticity robust estimator).

With this notation at hand, consider the problem of maximizing power in the following hypothesis testing problem:

$$H_0 : \ t_1 = 0, \ \ t_{12} \in \mathbb{R} \quad \text{against} \quad H_1 : \ t_1 \neq 0, \ \ t_{12} = 0. \tag{17}$$

A standard approach to deal with composite hypotheses is to consider a weighted average power criterion to choose among tests. With this approach, the hypothesis testing problem (17) becomes one of finding a powerful test for "H_0: the density of $\hat{t}$ is $f_t, t_1 = 0, t_{12} \in \mathbb{R}$" against the simple alternative "$H_{1,F}$: the density of $\hat{t}$ is $g = \int f_t dF(t)$",

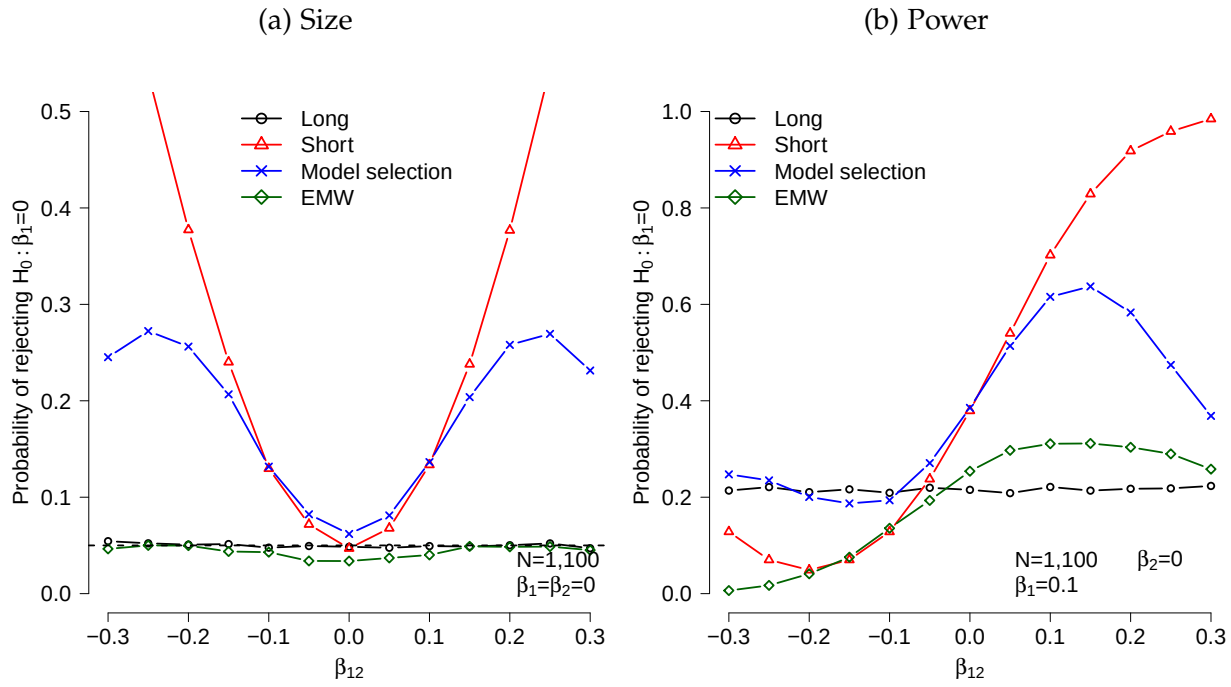
[17] This conclusion is specific to our simulation design. Based on a different data generating process, McCloskey (2017) finds local power gains relative to the long model.

where F is a user-specified weighting function. F describes the importance a researcher attaches to the ability of the test to reject under different alternatives.

Suppose that the null hypothesis is also replaced by a single hypothesis "$H_{0,\Lambda}$: the density of $\hat{t}$ is $\int f_t d\Lambda(t)$". The goal is then to find a least favorable distribution (LFD), Λ^{LF}, with the property that the size α Neyman-Pearson test of the simple hypothesis $H_{0,\Lambda^{LF}}$ against $H_{1,F}$ also yields a size α test of the composite null hypothesis H_0 against $H_{1,F}$ (e.g., Lehmann & Romano, 2005).

Because it is generally difficult to analytically determine Λ^{LF}, EMW suggest to instead focus on an approximate LFD, Λ^{ALF}, which yields a nearly optimal test for H_0 against $H_{1,F}$. The resulting test is then just a Neyman-Pearson test based on Λ^{ALF}.

Figure 5: Elliott et al. (2015)'s nearly optimal test controls for size and has power gains over running the full model for "intermediate" values of β_{12}



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 5a and 5b 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.

Figure 5 displays the results of applying the nearly optimal test in the context of our running example.[18] We can see that the test controls size for all values of β_{12} (see Figure 5a). Moreover, Figure 5b shows that the test exhibits more power than the standard t-test based on the long model when β_{12} is close to zero. However, these power gains come

[18]To improve the performance of their procedure, EMW suggest a switching rule that depends on $|\hat{t}_{12}|$ such that for large enough values of $|\hat{t}_{12}|$, one switches to regular hypothesis testing. Following their suggestion, we use 6 as the switching value.

at a cost: for certain values of β_{12}, the power can be much lower than the power of the t-test based on the long model (see also Figure E.4 in the Appendix)

5.2 Incorporating prior knowledge

Suppose that the researcher is certain that $\beta_{12} = \bar{\beta}_{12}$. In this case, she can obtain $\beta := (\beta_0, \beta_1, \beta_2)$ from a population regression of $Y - \bar{\beta}_{12}T_{12}$ on T_1, T_2 and a constant. Letting $X := (1, T_1, T_2)'$, the resulting regression population regression coefficients are given as

$$\beta = E\left(XX'\right)^{-1} E\left(X(Y - \beta_{12}T_{12})\right),$$

For example, if $\bar{\beta}_{12} = 0$, β is equal to the population regression coefficient in the short model,

$$\beta = E\left(XX'\right)^{-1} E\left(XY\right).$$

Of course, exact knowledge of β_{12} may be too strong of an assumption. Suppose instead that the researcher believes that β_{12} is smaller than c times the maximum of the two main effects that the experiment is powered to detect, $\bar{\beta}_1$ and $\bar{\beta}_2$.

Assumption 2. *For some $c \geq 0$, $|\beta_{12}| \leq c \cdot \max\left\{\left|\bar{\beta}_1\right|, \left|\bar{\beta}_2\right|\right\}$*

Assumption 2 implies that β_{12} lies in a compact interval,

$$\beta_{12} \in \left[-c \cdot \max\left\{\left|\bar{\beta}_1\right|, \left|\bar{\beta}_2\right|\right\}, c \cdot \max\left\{\left|\bar{\beta}_1\right|, \left|\bar{\beta}_2\right|\right\}\right] \equiv \left[\beta_{12}^l, \beta_{12}^u\right].$$

Under Assumption 2, point identification is generally lost and β lies in an identified set. Let $\beta(\beta_{12}) := (\beta_0(\beta_{12}), \beta_1(\beta_{12}), \beta_2(\beta_{12}))'$ denote the population regression coefficient from a regression of $Y - \beta_{12}T_{12}$ on X:

$$\begin{aligned}\beta(\beta_{12}) &:= E\left(XX'\right)^{-1} E\left(X(Y - \beta_{12}T_{12})\right) \\ &= E\left(XX'\right)^{-1} E\left(XY\right) - \beta_{12} E\left(XX'\right)^{-1} E\left(XT_{12}\right)\end{aligned}$$

$E\left(XX'\right)^{-1} E\left(XT_{12}\right) =: (\gamma_1, \gamma_2, \gamma_{12})'$ is the population regression coefficient from a regression of T_{12} on X. Independence of T_1 and T_2 (Assumption 1) implies hat $\gamma_1 = E(T_{12} \mid T_1 = 1) - E(T_{12} \mid T_1 = 0)$ and $\gamma_2 = E(T_{12} \mid T_2 = 1) - E(T_{12} \mid T_2 = 0)$ both of which are positive. Consequently, the identified set for β_t, $t \in \{1, 2\}$, is given by

$$\beta_t \in \left\{\beta_t(\beta_{12}), \beta_{12} \in [\beta_{12}^l, \beta_{12}^u]\right\} = \left[\beta_t(\beta_{12}^u), \beta_t(\beta_{12}^l)\right] =: \left[\beta_t^l, \beta_t^u\right].$$

The lower bound β_t^l can be estimated from an OLS regression of $Y - \beta_{12}^u T_{12}$ on X. Similarly, the upper bound β_t^u can be obtained from an OLS regression of $Y - \beta_{12}^l T_{12}$ on X. Under standard conditions, the OLS estimators $\hat{\beta}_t^l$ and $\hat{\beta}_t^u$ are asymptotically normal and the asymptotic variances $Avar\left(\hat{\beta}_t^l\right)$ and $Avar\left(\hat{\beta}_t^u\right)$ can be estimated consistently. We therefore apply the approach of Imbens & Manski (2004) and Stoye (2009) to construct confidence intervals for β_t:[19]

$$CI_{1-\alpha} = \left[\hat{\beta}_t^l - c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^l\right)}{N}},\ \hat{\beta}_t^u + c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^u\right)}{N}}\right], \tag{18}$$

where the critical value c_{IM} solves

$$\Phi\left(c_{IM} + \sqrt{N} \cdot \frac{\hat{\beta}_t^u - \hat{\beta}_t^l}{\sqrt{\max\left(\widehat{Avar}\left(\hat{\beta}_t^l\right), \widehat{Avar}\left(\hat{\beta}_t^l\right)\right)}}\right) - \Phi\left(-c_{IM}\right) = 1 - \alpha.$$

Imbens & Manski (2004) and Stoye (2009) show that this is a valid confidence interval for β_t. Finally, while we focus on Assumption 2 here, the construction of the confidence interval discussed above can directly be applied whenever the researcher believes that β_{12} lies in some compact interval.

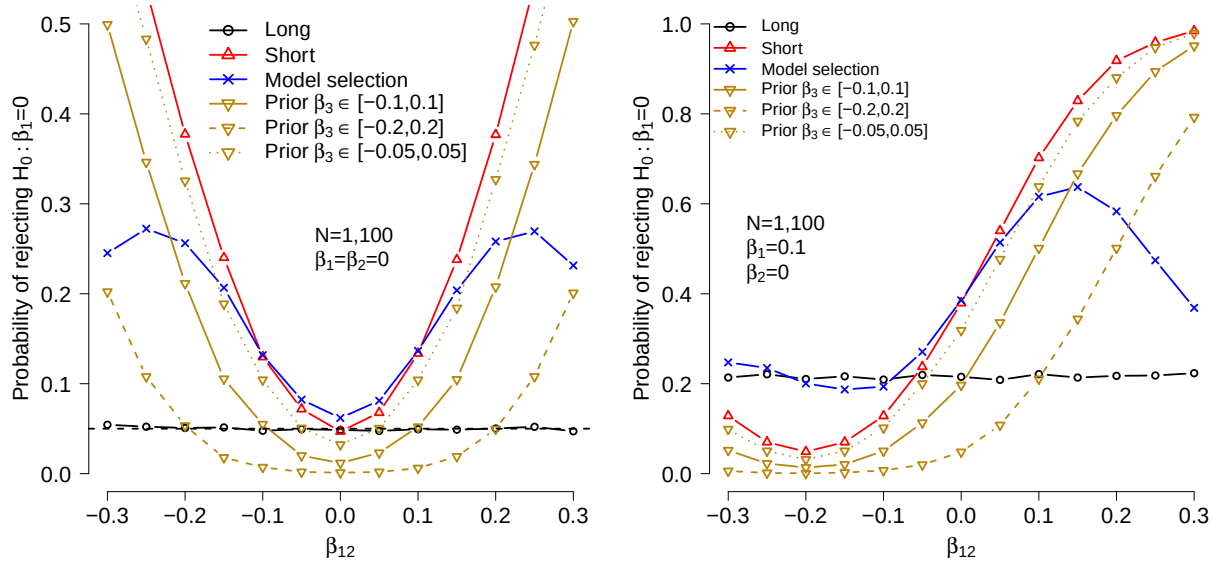
In Figure 6, we report the rejection probabilities of a test that rejects if zero is not in the confidence interval (18). For the purpose of illustration, we consider three different cases: $\beta_{12} \in [-0.1, 0.1]$, $\beta_{12} \in [-0.2, 0.2]$, and $\beta_{12} \in [-0.05, 0.05]$. When $\max\{|\bar{\beta}_1|, |\bar{\beta}_2|\} = 0.1$, these intervals correspond to specifying $c = 1$, $c = 2$ and $c = 0.5$ in Assumption 2. Our results suggest that imposing prior knowledge can improve power relative to the long regression model, while controlling size when this prior knowledge is in fact correct. However, this method may exhibit substantial size distortions when the prior knowledge is incorrect. The width of the identified set is increasing in c and, consequently, the potential power gains are decreasing in c.

[19] By construction, the upper bound is always weakly larger than the lower bound. Hence Lemma 3 in Stoye (2009) justifies the procedure in Imbens & Manski (2004).

Figure 6: Prior knowledge of β_{12} can lead to power gains if the prior knowledge is correct but yields incorrect inferences if it is not

(a) Size

(b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 6a and 6b is $\alpha = 0.05$.

6 Should we run experiments with factorial designs?

The discussion so far begs the question: should we run experiments with factorial designs in the first place? A natural alternative is to leave the "interaction cell" empty (i.e., to set $N_4 = 0$) and to re-assign those subjects to the other cells. This implies that $\lim_{N\to\infty} N_4/N = P(T_1 = 1, T_2 = 1) = 0$ such that T_1 and T_2 are not independent if $P(T_1 = 1)$ and $P(T_2 = 1)$ are both non-zero, which is the case of interest in this paper.[20] We therefore consider the following weaker version of the random assignment assumption.

Assumption 3. $(Y_{0,0}, Y_{1,0}, Y_{0,1}, Y_{1,1}) \perp\!\!\!\perp (T_1, T_2)$

Under Assumption 3, the ATEs defined in Section 2.1 and the interaction effect are

[20] Independence of T_1 and T_2 means that $P(T_1 = 1, T_2 = 1) = P(T_1 = 1)P(T_2 = 1)$.

identified. In Appendix C.2, we show that if $P(T_1 = 1, T_2 = 1) = 0$,

$$\begin{aligned}
\beta_1^s &= E\left(Y \mid T_1 = 1, T_2 = 0\right) - E\left(Y \mid T_1 = 0, T_2 = 0\right) \\
&= E\left(Y_{1,0} - Y_{0,0}\right), \\
\beta_2^s &= E\left(Y \mid T_1 = 0, T_2 = 1\right) - E\left(Y \mid T_1 = 0, T_2 = 0\right) \\
&= E\left(Y_{0,1} - Y_{0,0}\right).
\end{aligned}$$

Thus, the OLS estimators based on the short model are consistent for the main effects.

Consider next the power implications of leaving the interaction cell empty. Specifically, consider an experiment where the N_4 individuals in the interaction cell are equally distributed to the other cells:

		T_1	
		No	*Yes*
T_2	*No*	$N_1 + \frac{N_4}{3}$	$N_2 + \frac{N_4}{3}$
	Yes	$N_3 + \frac{N_4}{3}$	0

In what follows, we focus on β_1^s. The analysis for β_2^s is symmetric and omitted. Let $\hat{\beta}_1^{nf}$ denote the OLS estimator based on the short model, leaving the interaction cell empty. Under the assumptions of Section 2.3, the variance of $\hat{\beta}_1^{nf}$ is given by

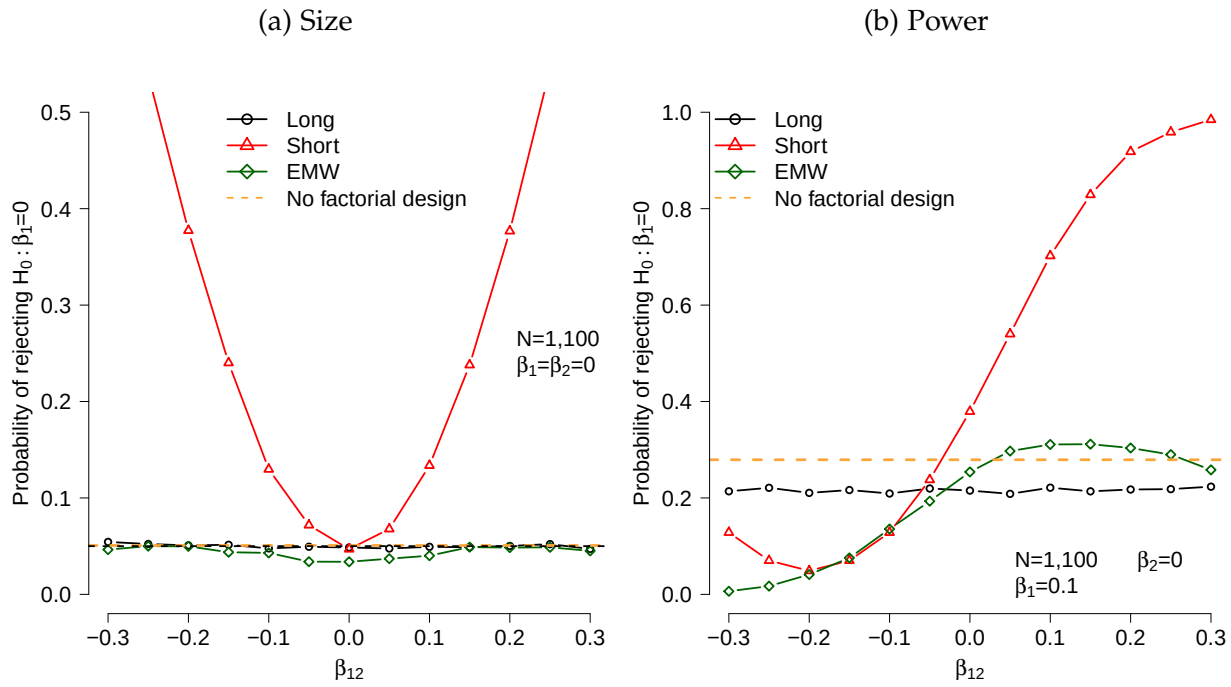
$$Var\left(\hat{\beta}_1^{nf}\right) = \sigma^2 \frac{\left(N_1 + \frac{N_4}{3}\right) + \left(N_2 + \frac{N_4}{3}\right)}{\left(N_1 + \frac{N_4}{3}\right)\left(N_2 + \frac{N_4}{3}\right)}$$

A comparison with the variance of $\hat{\beta}_1$ shows that $Var\left(\hat{\beta}_1^{nf}\right) < Var\left(\hat{\beta}_1\right)$.[21] Thus, by the same reasoning as in Section 2.3, leaving the interaction cell empty leads to power improvements for testing hypotheses about β_1 relative to the long regression model.

Figure 7 presents the results based on our running example. As expected, leaving the interaction cell empty yields tests that control size for all values of β_{12}. Moreover, among the approaches that achieve uniform size control, leaving the interaction cell empty yields the highest finite sample power for a wide range of relevant values of β_{12}. Thus, if one is not interested in interaction effects, we recommend avoiding factorial designs and leaving the interaction cell empty as done for instance in the design employed by Muralidharan & Sundararaman (2011).

[21] Note that $Var\left(\hat{\beta}_1\right) - Var\left(\hat{\beta}_1^{nf}\right) = \sigma^2 \frac{N_4(3N_1^2 + N_1N_4 + 3N_2^2 + N_2N_4)}{N_1N_2(3N_1 + N_4)(3N_2 + N_4)} \geq 0$.

Figure 7: Leaving the interaction cell empty increases power for most values of β_{12} relative to other alternatives



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 7a and 7b is $\alpha = 0.05$. When there is no factorial design, we assume the N_4 individuals in the interaction cell are equally distributed to the other cells.

7 Factorial designs with more than two treatments

So far, our theoretical discussion has focused on 2×2 factorial designs. Here we briefly discuss designs with more than two treatments. To fix ideas, consider a 2×2×2 factorial design with three different treatments T_1, T_2 and T_3. The long and the short regression model are

$$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \beta_{12} T_1 T_2 + \beta_{13} T_1 T_3 + \beta_{23} T_2 T_3 + \beta_{123} T_1 T_2 T_3 + \epsilon \quad (19)$$

$$Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \beta_3^s T_3 + \varepsilon^s \quad (20)$$

The long model (19) now features four interaction terms, three two-way interactions, T_1T_2, T_2T_3 and T_1T_3 and one three-way interaction, $T_1T_2T_3$.

The theoretical analysis of Section 2 straightforwardly extends to this case (and more general cases as well). In particular, the estimators based on (19) are consistent for the main and interaction effects, whereas the estimators based on (20) are inconsistent for the main effects unless the interactions are zero. Furthermore, the t-tests based on (20)

are more powerful than the corresponding t-tests based on (19) when the interactions are zero, but may suffer from size distortions when they are not. The estimands based on the short model become substantially harder to interpret when there are many treatments.

Conceptually, all the econometric approaches discussed in Section 5 can be extended beyond 2×2 settings. However, the nearly optimal tests become computationally prohibitive when there are many interactions (i.e., many nuisance parameters) and cannot be recommended for complicated factorial designs. By contrast, the Bonferroni approach may still be computationally tractable and provides an alternative to the long regression model. In view of the simulation results in Section 5 and in McCloskey (2017), we expect the relative finite sample performance of the Bonferroni approach and the long regression model to be design-specific. Incorporating prior knowledge, while computationally feasible even in very high-dimensional cross-cut designs, can be problematic in practice because this approach requires reliable prior knowledge on potentially very many interactions. As in 2×2 designs, leaving the interaction cells empty at the design stage provides power improvements over the long model.

Figure 8b provides simulation evidence based on the following data generating process:

$$Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \beta_{12} T_1 T_{2i} + \beta_{13} T_{1i} T_{3i} + \beta_{23i} T_{2i} T_{3i} + \beta_{123} T_{1i} T_{2i} T_{3i} + \epsilon_i.$$

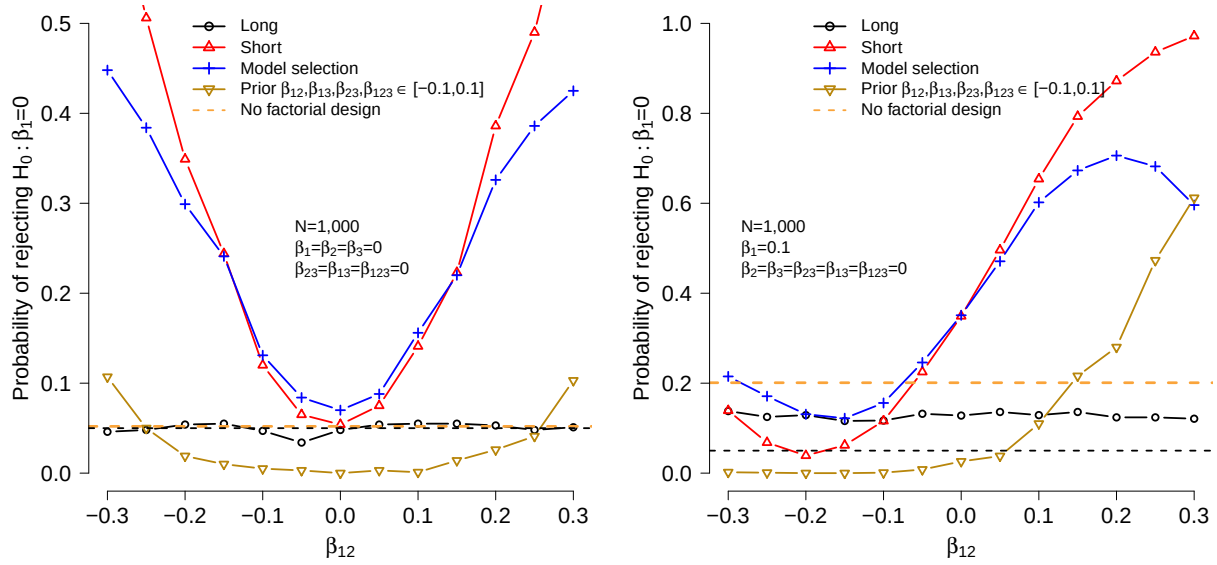
Here $\varepsilon_i \sim N(0,1)$ and T_{1i}, T_{2i}, and T_{3i} are mutually independent and randomly assigned treatments with $P(T_{1i} = 1) = P(T_{2i} = 1) = P(T_{3i} = 1) = 0.5$. The simulation results are based on 10,000 repetitions and show the bias and the empirical rejection rates under the null (size) and the alternative hypothesis (power). To keep the simulations computationally tractable, we do not report results for nearly optimal tests and the Bonferroni approach.[22]

The results in Figure 8b corroborate our findings for 2×2 designs. Ignoring the interactions and using data-driven model selection generally yields incorrect inferences for the main effects. Imposing prior knowledge improves the finite sample power over some parts of the parameters space, but leads to size distortions if the prior knowledge is not correct. Leaving the interaction cell empty yields power gains relative to the long model while controlling size uniformly across all values of the interactions. This approach is recommended whenever the interaction effects are not of primary interest.

[22]While the Bonferroni approach may remain tractable for a single study, evaluating its performance in Monte Carlo simulations is computationally prohibitive.

Figure 8: Size and power in a 2x2x2 design

(a) Size (b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 8a and 8b is $\alpha = 0.05$*. When there is no factorial design, we assume individuals in the interaction cell are equally distributed to the other cells.*

8 Conclusion

If there are no interaction effects between different treatments, cross-cut designs provide a cost-effective way to increase the effective sample size and statistical power. Specifically, *t*-tests based the short regression model, which only includes the main treatment indicators and omits the interactions, are more powerful than *t*-tests based on the long regression model, which also includes the interactions. However, if the interaction effects are not zero, the estimators based on the short regression model are inconsistent and the corresponding *t*-tests may exhibit substantial size distortions. Hence, there is a fundamental trade-off between the short model, which improves power but yields incorrect inferences unless the interaction effects are zero, and the long model, which yields correct inferences irrespective of the true value of the interaction effects.

Can we improve power relative to the long regression model, while at the same time controlling size? A natural idea is to engage in data-driven model selection, where the decision of whether to include the interactions into the regression model is based on pre-testing. Consistent with a large body of theoretical work, we document that *t*-

statistic-based inference after model selection is invalid for practically relevant values of the interaction effects. Hence, this approach cannot be recommended.

We therefore explore the performance of two state-of-the-art inference procedures that achieve size control irrespective of the true value of the interaction effects: model selection combined with a Bonferroni correction step (McCloskey, 2017) and a nearly optimal test which targets power towards pre-specified and a priori likely values of the interaction effects (Elliott et al., 2015). In our application-based simulations, we find that only the nearly optimal test yields power improvements over the t-test based on the long model for economically relevant values of the interactions. By construction, these power improvements are achieved in the vicinity of the a priori likely values and come at the cost of potentially very low power for other values of the interactions. As an alternative, we explore a partial identification approach based on prior knowledge about the value of the interaction effect. Specifically, we assume that the interaction effect is no larger than a pre-specified constant times the maximum main effect that the experiment is powered to detect. When the prior knowledge is correct, this approach can achieve substantial power improvements, but it fails to control size whenever the prior information is incorrect.

All three approaches are suitable for existing cross cut designs. For the design of new experiments, a natural alternative is to leave the interaction cells empty. This simple design-based approach naturally controls size and, in our simulations, yields notable power improvements relative to the long model. This is the approach we recommend for policy experiments where a "business as usual" counterfactual is especially important.

In other cases, researchers may be willing to make the case that the parameter of interest for their studies is not the main treatment effect relative to a "business as usual" counterfactual but the composite effect that includes a weighted average of interactions with the other treatments in the experiment. In such cases, we recommend that this point be clearly committed to in a pre-analysis plan and that the write up of the results be explicit about how the presented treatment effects should be interpreted. This will increase transparency in the reporting of results (especially for complex designs with several interactions) and allow readers to assess the extent to which the other treatments can reasonably be interpreted as "background" omitted variables.

References

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., & Tobias, J. (2012, June). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, *102*(4), 1206-40. doi: 10.1257/aer.102.4.1206

Allcott, H., & Taubinsky, D. (2015, August). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, *105*(8), 2501-38. doi: 10.1257/aer.20131564

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008, may). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618. doi: 10.1111/j.1468-0262.2008.00848.x

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653.

Andrews, D. W. K., & Guggenberger, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica*, *77*(3), 721-762.

Ashraf, N., Berry, J., & Shapiro, J. M. (2010, December). Can higher prices stimulate product use? evidence from a field experiment in zambia. *American Economic Review*, *100*(5), 2383-2413. doi: 10.1257/aer.100.5.2383

Athey, S., & Imbens, G. W. (2017). Chapter 3 - the econometrics of randomized experimentsa. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments* (Vol. 1, p. 73 - 140). North-Holland. doi: https://doi.org/10.1016/bs.hefe.2016.10.003

Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). What drives taxi drivers? a field experiment on fraud in a market for credence goods. *Review of Economic Studies*, *80*(3), 876–891.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235-1264.

Banerjee, A., & Duflo, E. (2005). Chapter 7 growth theory through the lens of development economics. In P. Aghion & S. N. Durlauf (Eds.), (Vol. 1, p. 473 - 552). Elsevier.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, *125*(1), 263-306. doi: 10.1162/qjec.2010.125.1.263

Blattman, C., Jamison, J. C., & Sheridan, M. (2017, April). Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia. *American Economic Review*, *107*(4), 1165-1206. doi: 10.1257/aer.20150503

Brown, J., Hossain, T., & Morgan, J. (2010). Shrouded attributes and information suppression: Evidence from the field. *The Quarterly Journal of Economics*, *125*(2), 859–876.

Bruhn, M., & McKenzie, D. (2009, October). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*(4), 200-232. doi: 10.1257/app.1.4.200

Carneiro, P., Lee, S., & Wilhelm, D. (2017). *Optimal data collection for randomized control trials.* cemmap working paper CWP15/17.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs*. John Wiley & Sons.

Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, *125*(1), 1-45. doi: 10.1162/qjec.2010.125.1.1

Cohen, J., Dupas, P., & Schaner, S. (2015). Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial. *American Economic Review*, *105*(2), 609–45.

DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, *84*(1), 143–181.

Duflo, E., Dupas, P., & Kremer, M. (2008, July). *Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya.* Retrieved from http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1239047988859/5995659-1239051886394/5996104-1246378480717/Dupas_ETP_07.21.08.pdf (Working paper)

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739-74. doi: 10.1257/aer.101.5.1739

Duflo, E., Dupas, P., & Kremer, M. (2015a, September). Education, hiv, and early fertility: Experimental evidence from Kenya. *American Economic Review*, *105*(9), 2757-97.

Duflo, E., Dupas, P., & Kremer, M. (2015b). School governance, teacher incentives, and pupil-teacher experiemental evidence from Kenyan primary schools. *Journal of Public Economics*, *123*, 92-110.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, *4*, 3895–3962.

Elliott, G., Müller, U. K., & Watson, M. W. (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica*, *83*(2), 771–811.

Eriksson, S., & Rooth, D.-O. (2014, March). Do employers use unemployment as a sorting criterion when hiring? evidence from a field experiment. *American Economic Review*, *104*(3), 1014-39. doi: 10.1257/aer.104.3.1014

Fischer, G. (2013). Contract structure, risk-sharing, and investment choice. *Econometrica*, *81*(3), 883–939.

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2008). Racial preferences in dating. *The Review of Economic Studies*, *75*(1), 117–132.

Flory, J. A., Leibbrandt, A., & List, J. A. (2014). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, *82*(1), 122–155.

Gelman, A. (2018, Mar). *You need 16 times the sample size to estimate an interaction than to estimate a main effect.* Retrieved from https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Gerber, A., & Green, D. (2012). *Field experiments: Design, analysis, and interpretation*. W. W. Norton.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D. (2018, May). *Educator Incentives and Educational Triage in Rural Primary Schools* (IZA Discussion Papers No. 11516).

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637–1664.

Haushofer, J., & Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya. *The Quarterly Journal of Economics*, *131*(4), 1973–2042.

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*(6), 1845–1857. doi: 10.1111/j.1468-0262.2004.00555.x

Jakiela, P., & Ozier, O. (2015). Does africa need a rotten kin theorem? experimental evidence from village economies. *The Review of Economic Studies*, *83*(1), 231–268.

Karlan, D., & List, J. A. (2007, December). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774-1793. doi: 10.1257/aer.97.5.1774

Karlan, D., Osei, R., Osei-Akoto, I., & Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, *129*(2), 597–652.

Karlan, D., & Zinman, J. (2008, June). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, *98*(3), 1040-68. doi: 10.1257/aer.98.3.1040

Karlan, D., & Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, *77*(6), 1993–2008.

Kaur, S., Kremer, M., & Mullainathan, S. (2015). Self-control at work. *Journal of Political Economy*, *123*(6), 1227–1277.

Kendall, C., Nannicini, T., & Trebbi, F. (2015, January). How do voters respond to information? evidence from a randomized campaign. *American Economic Review*, *105*(1), 322-53. doi: 10.1257/aer.20131063

Kerwin, J. T., & Thornton, R. L. (2017). *Making the grade: The trade-off between efficiency and effectiveness in improving student learning* (Working Paper). University of Minnesota.

Khan, A. Q., Khwaja, A. I., & Olken, B. A. (2015). Tax farming redux: Experimental evidence on performance pay for tax collectors. *The Quarterly Journal of Economics*, *131*(1), 219–271.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in Denmark. *Econometrica*, *79*(3), 651-692. doi: 10.3982/ECTA9113

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *The American Economic Review*, *93*(2), pp. 102-106.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21-59.

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 2554–2591.

Leeb, H., & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(02), 338–376.

Leeb, H., & Pötscher, B. M. (2017). Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values. In S. E. Ahmed (Ed.), *Big and complex data analysis: Methodologies and applications* (pp. 69–82). Cham: Springer International Publishing.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439.

List, J. A., Shaikh, A. M., & Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1–21.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019, 04). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *The Quarterly Journal of Economics*, *134*(3), 1627-1673. doi: 10.1093/qje/qjz010

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*.

McCloskey, A. (2019). Asymptotically uniform tests after consistent model selection in the linear regression model. *Journal of Business & Economic Statistics*, *0*(0), 1-35. doi: 10.1080/07350015.2019.1592754

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, *115*(2), 200-249. doi: 10.1086/517935

Pallais, A., & Sands, E. G. (2016). Why the referential treatment? evidence from field experiments on referrals. *Journal of Political Economy*, *124*(6), 1793–1828.

Ray, D. (1998). *Development economics*. Princeton University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, *77*(4), 1299–1315. doi: 10.3982/ECTA7347

Thornton, R. L. (2008, December). The demand for, and impact of, learning HIV status. *American Economic Review*, *98*(5), 1829-63. doi: 10.1257/aer.98.5.1829

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

Young, A. (in press). Channelling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*.

A Papers with factorial designs published in Top-5 economics journals

Table A.1: Papers with factorial designs published between 2006 and 2017 in top-5 economics journals sorted by citation count (as of April of 2018)

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Olken (2007)	Monitoring Corruption: Evidence from a Field Experiment in Indonesia	JPE	2007	1529	3	2	0	Yes
Banerjee et al. (2007)	Remedying Education: Evidence from Two Randomized Experiments in India	QJE	2007	1213	2	1	0	Yes
Duflo et al. (2011)	Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya	AER	2011	787	3	4	0	Yes
Kleven et al. (2011)	Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark	ECMA	2011	776	2	1	0	No

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Karlan et al. (2014)	Agricultural Decisions after Relaxing Credit and Risk Constraints	QJE	2014	612	2	1	1	
Bertrand et al. (2010)	What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment	QJE	2010	522	14	85	0	Yes
Karlan & List (2007)	Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment	AER	2007	506	7	28	0	Yes
Thornton (2008)	The Demand for, and Impact of, Learning HIV Status	AER	2008	453	2	1	0	Yes
Haushofer & Shapiro (2016)	The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya	QJE	2016	393	6	8	3	Yes

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Alatas et al. (2012)	Targeting the Poor: Evidence from a Field Experiment in Indonesia	AER	2012	330	4	16	0	Yes
Karlan & Zinman (2008)	Credit Elasticities in Less-Developed Economies: Implications for Microfinance	AER	2008	311	3	2	0	Yes
Duflo et al. (2015a)	Education, HIV, and Early Fertility: Experimental Evidence from Kenya	AER	2015	282	3	3	1	Yes
Andreoni et al. (2017)	Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving	JPE	2017	270	2	1	1	
Jakiela & Ozier (2015)	Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies	ReStud	2016	245	3	6	6	

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Eriksson & Rooth (2014)	Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment	AER	2014	238	34	71680	0	Yes
Allcott & Taubinsky (2015)	Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market	AER	2015	237	2	1	1	
Flory et al. (2014)	Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions	ReStud	2015	204	2	1	1	
Brown et al. (2010)	Shrouded Attributes and Information Suppression: Evidence from the Field	QJE	2010	189	3	6	6	
DellaVigna et al. (2016)	Voting to Tell Others	ReStud	2017	169	4	15	0	Yes
Fischer (2013)	Contract Structure, Risk-Sharing, and Investment Choice	ECMA	2013	162	2	1	1	

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Kaur et al. (2015)	Self-Control at Work	JPE	2015	154	2	1	1	
Cohen et al. (2015)	Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial	AER	2015	151	3	7	7	
Blattman et al. (2017)	Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia	AER	2017	135	2	1	1	
Khan et al. (2015)	Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors	QJE	2016	133	6	8	0	Yes
Balafoutas et al. (2013)	What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods	ReStud	2013	126	5	6	0	Yes

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available
Kendall et al. (2015)	How Do Voters Respond to Information? Evidence from a Randomized Campaign	AER	2015	116	3	2	2	
Pallais & Sands (2016)	Why the Referential Treatment? Evidence from Field Experiments on Referrals	JPE	2016	85	3	12	0	No

Note: This table provides relevant information from the most cited articles with factorial designs. Citation counts are from Google Scholar on July 4th of 2019. Treatments is the number of different treatments in the paper. "Interactions in Design" is the number of interactions in the experimental design. "Interactions Included" is the number of interactions included in the main specification of the paper. Data available, refers to whether the data is publicly available or not.

A.1 Monitoring Corruption: Evidence from a Field Experiment in Indonesia

Olken (2007) analyze an experiment with a factorial design in which several villages are randomized into three interventions: i) Increasing the probability of external audits ("audits"), ii) increasing participation in accountability meetings ("invitations"), and iii) allowing villagers to provide anonymous comments ("invitations plus comments"). As the paper notes "randomization into the "invitations" and "invitations plus comments" treatments was independent of randomization into the "audits" treatment". Figure A.1 — taken from the published version of the paper — shows the details of the randomization design. The estimating equation does not include the interaction term. The paper does not mention that the treatment effects for each treatment is the weighted average over the other treatments. For example, the audit results are presented as "The results show that the audits had a substantial, and statistically significant, negative effect on the percentage of expenditures that could not be accounted for". The invitation results are presented as "The results in column 1 suggest that neither the invitations treatment nor the invitations plus comment forms treatment had a significant effect on the total number of problems discussed at the meeting". The paper does not contain a table in the main text, nor in the Appendix where the full model is estimated. We re-estimate the main results in the paper (Column 3 of Table 4 and Table 11) using the long model.

Figure A.1: Factorial design in Olken (2007)

TABLE 1
NUMBER OF VILLAGES IN EACH TREATMENT CATEGORY

	Control	Invitations	Invitations Plus Comment Forms	Total
Control	114	105	106	325
Audit	93	94	96	283
Total	207	199	202	608

NOTE.—Tabulations are taken from results of the randomization. Each subdistrict faced a 48 percent chance of being randomized into the audit treatment. Each village faced a 33 percent chance of being randomized into the invitations treatment and a 33 percent chance of being randomized into the invitations plus comment forms treatment. The randomization into audits was independent of the randomization into invitations or invitations plus comment forms.

Note: Table 1 from Olken (2007).

A.2 Remedying Education: Evidence from Two Randomized Experiments in India

Banerjee et al. (2007) analyze an experiment with a factorial design in which several schools are assigned, over a three year period, to a remedial education program (Balsakhi) or a Computer-Assisted Learning (CAL) program. The details of the factorial design are summarized in Figure A.2, taken from the published version of the paper. Since the factorial design only took place in fourth grade schools in Vadodara, we re-estimate the results of the paper that focus on this population. We re-estimate the results in Table 3 (Column 4, Panel D, Year 2) of the original paper and the results in Table 4 (Column 4, Panels A and B, Year 2) of the original paper.

The paper does not mention that the treatment effects for each treatment is the weighted average over the other treatments. The paper does present the interactions after the main tables. Explicitly, "Panel B of Table IV compares the Balsakhi and the CAL effects and examines their interactions in year 2 (2002-2003) when they were implemented at the same time using a stratified design. When the two programs are considered in isolation, the CAL has a larger effect on math test scores than the Balsakhi Program (although this difference is not significant) and a smaller effect on overall test scores (although, again, the difference is not significant). The programs appear to have no interaction with each other: the coefficients on the interaction on the math and overall test score are negative and insignificant."

Figure A.2: Factorial design in Banerjee et al. (2007)

TABLE I
SAMPLE DESIGN AND TIME LINE

	Year 1 (2001–2002)		Year 2 (2002–2003)		Year 3 (2003–2004)	
	Grade 3	Grade 4	Grade 3	Grade 4	Grade 3	Grade 4
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Vadodara						
Balsakhi						
Group A (5,264 students in 49 schools in year 1; 6,071 students in 61 schools in year 2)	Balsakhi	No balsakhi	No balsakhi	Balsakhi	No balsakhi	No balsakhi
Group B (4,934 students in 49 schools in year 1; 6,344 students in 61 schools in year 2)	No balsakhi	Balsakhi	Balsakhi	No balsakhi	No balsakhi	No balsakhi
Computer-Assisted Learning (CAL)						
Group A1B1 (2,850 students in 55 schools in year 2; 2,814 students in 55 schools in year 3)	No CAL	No CAL	No CAL	CAL	No CAL	No CAL
Group A2B2 (3,095 students in 56 schools in year 2; 3,131 students in 56 schools in year 3)	No CAL	No CAL	No CAL	No Cal	No CAL	CAL
Panel B: Mumbai						
Balsakhi						
Group C (2,592 students in 32 schools in year 1; 5,755 students in 38 schools in year 2)	Balsakhi	No balsakhi	No balsakhi	Balsakhi	No balsakhi	No balsakhi
Group D (2,182 students in 35 schools year 1; 4,990 students in 39 schools in year 2)	No balsakhi	No balsakhi	Balsakhi	No balsakhi	No balsakhi	No balsakhi

Notes: This table displays the assignment to schools in various treatment groups in the three years of the evaluation.
Group A1B1 and A2B2 were constituted by randomly assigning half the schools in Group A and half the schools in Group B to the Group A1B1 and the remaining schools to the Group A2B2.
Schools assigned to Group A (resp. B) in 2001–2002 remained in Group A (resp. B) in 2002–2003. Twelve new schools were brought in the study and assigned randomly to Groups A and B.
Schools assigned to Group C (resp. D) in 2001–2002 remained in Group C (resp. D) in 2002–2003. Ten new schools were brought in the study and assigned randomly to Groups C and D.

Note: Table 1 from Banerjee et al. (2007).

A.3 Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya

The evaluation featured a factorial design with three treatments: Extra contract teacher; school based management; and tracking (i.e., splitting classes by ability). Figure A.3 taken from Duflo et al. (2008) working paper has details of the experimental design. The published version of the paper does not mention the school based management treatment. The long model is not presented in any table in the paper, nor in the appendix. The paper does not mention that the treatment effects for each treatment is the weighted average over the other treatments. We re-estimate the results of Table IV (Panel A, Column 1) in Duflo et al. (2011) using the long model.[23].

[23] Duflo et al. (2015b) only includes the sample of schools with an extra contact teacher and school based management (dropping the sample of schools with tracking) and study the interactions between these two treatments.

Figure A.3: Factorial design in Duflo et al. (2011) and Duflo et al. (2015b)

Figure 1
Experimental Design: The Extra-Teacher Project

Group	# Schools	Class Size	Peer Grouping	Training on School-Based Management of Teachers (SBM)	Teacher Employer	# Classes
Non-ETP Schools (Comparison)	70	Normal	Unchanged	No	Government	88
Non-Tracked Schools	70	Reduced	Random	No	Government	41
					School Committee	35
				Yes	Government	42
					School Committee	35
Tracked Schools	70	Reduced	Tracking by Initial Achieve-ment	No	Government	41
					School Committee	35
				Yes	Government	41
					School Committee	35

Note: Table 1 from Duflo et al. (2008).

A.4 Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark

Kleven et al. (2011) analyze a tax enforcement field experiment in Denmark. The experiment features a factorial with two independent treatments. The first is a random audit and the second are threat-of-audit letters. The data is not available online. After the main tables, Table VI analyzes the effects of one treatment (information letters) conditional on the other treatment (audit), from which they conclude that "letter effects are roughly the same in the 0% and 100% audit groups." The paper does not mention that the treatment effects for each treatment is the weighted average over the other treatments.

A.5 Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment

Karlan & List (2007) analyze a field experiment with a factorial design in which letters requesting donations are randomized across three dimensions: matching ratio, maximum matching quantity, and a donation suggestion. As the paper states, they "use several treatments and sub-treatments that span the range of design parameters that fundraisers are most likely to utilize". Regarding interactions, the paper further explains that "In terms of the other treatment variables, the figures suggest that neither the match threshold nor the example amount had a meaningful influence on behavior... Although our estimates are imprecisely measured, after interacting the match ratios and threshold amounts fully, we do not find systematic patterns for the interaction effects." The long model is not presented in any table in the paper, nor in the appendix. The paper does not mention that the treatment effects for each treatment is the weighted average over the other treatments. We re-estimate the results of Table 4 (Panel A, Column 1 and 2) in Karlan & List (2007) including all possible interactions.

A.6 What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment

Bertrand et al. (2010) analyze a mail field experiment in South Africa implemented by a consumer lender that randomized advertising content, loan price, and loan offer deadlines simultaneously. The experiment has a factorial design in which 14 features of the letter (and offer) are independently randomized. The paper does not include interactions terms, and is explicit about this: "We ignore interaction terms, given that we did not have any strong priors on the existence of interaction effects across treatments. Below, we motivate and detail our treatment design and priors on the main effects and groups of main effects." However, the paper does not explicitly mention that the treatment effects for each treatment is the weighted average over the other treatments. We replicate the paper including all possible two way interactions, but there are higher order interactions implied by the factorial design. We re-estimate the main results of the paper (Table 3, Column 1) using a linear probability model instead of a probit model. However, we only include two-way interactions in our re-estimation.

A.7 The Demand for, and Impact of, Learning HIV Status

Thornton (2008) analyzes an experiment in which individuals in rural Malawi are randomly assigned monetary incentives to learn their HIV results after being tested. The location of the HIV results centers was also randomly assigned (and hence the distance to the nearest center). After the main results (Table 4) the paper explores the interactions between the two treatments. Explicitly, the paper states: "Monetary incentives were also especially important for those living farther from the VCT center: for those living over 1.5 kilometers from the HIV results center, there was an additional impact of receiving an incentive, increasing attendance by 3.7 percentage points, although the difference is not statistically significant (Table 5, column 4). This effect can also be seen in Figure 4, panel B, which graphs the impact of distance on attendance among those receiving any incentive and those receiving no incentive." However, the paper does not mention that the treatment effects in the main tables (e.g., Table 4) are the weighted average over the other treatments. We re-estimate the results in Table 4 (Column 4) including the interaction between the incentives and the distance to the testing center.

A.8 Targeting the Poor: Evidence from a Field Experiment in Indonesia

Alatas et al. (2012) analyze an experiment in Indonesia, in which villages are randomly assigned to different targeting methods to distribute a cash transfer program. In some villages the targeting is done using a proxy-means test, in some the targeting is done by the community, and in some is a hybrid of both. In "community" and "hybrid" villages the treatments had several variations: In some villages the meetings took place during the day, in others at night. In some, the "elite" of the village took the decision, in some it was the whole community. In some, the 10 poorest households were primed by the meeting facilitator, in some there was no priming. Explicitly, the paper states "We designed several subtreatments in order to test three hypotheses about why the results from the community process might differ from those that resulted from the PMT treatment: elite capture, community effort, and within-community heterogeneity in preferences." Figure A.4 taken from Alatas et al. (2012) has details of the experimental design. However, the paper does not mention that the treatment effects in the main tables (e.g., Tables 3 and 4) are the weighted average over the subtreatments. Explicitly, the paper states "the PMT treatment is the omitted category, so β_1 and β_2 are interpretable as the impact of the community and the hybrid treatments relative to the PMT treatment". After the main results, Tables 7 explores the "elite" subtreatment. We re-estimate the results in Table 3

(Column 1) including all possible interactions.

Figure A.4: Factorial design in Alatas et al. (2012)

TABLE 1—RANDOMIZATION DESIGN

Community/hybrid subtreatments			Main treatments		
			Community	Hybrid	PMT
Elite	10 poorest first	Day	24	23	
		Night	26	32	
	No 10 poorest first	Day	29	20	
		Night	29	34	
Whole community	10 poorest first	Day	29	28	
		Night	29	23	
	No 10 poorest first	Day	28	33	
		Night	20	24	
		Total	214	217	209

Notes: This table shows the results of the randomization. Each cell reports the number of subvillages randomized to each combination of treatments. Note that the randomization of subvillages into main treatments was stratified to be balanced in each of 51 strata. The randomization of community and hybrid subvillages into each subtreatment (elite or full community, 10 poorest prompting or no 10 poorest prompting, and day or night) was conducted independently for each subtreatment, and each randomization was stratified by main treatment and geographic stratum.

Note: Table 1 from Alatas et al. (2012).

A.9 Credit Elasticities in Less-Developed Economies: Implications for Microfinance

Karlan & Zinman (2008) analyze an experiment in South Africa in which a lender sent out direct mail offers to over 50,000 former clients. The letters had a randomly assigned offer interest rate and in some cases a randomly assigned, nonbinding example maturity (four, six, or twelve months). In addition, each client was assigned a randomly selected a "contract rate" that was weakly less than the offer rate received by mail and revealed only after the borrower had accepted the solicitation and applied for a loan. We do not

study the re-randomization of the interest rate.[24] However, the paper does not mention that the treatment effects in the main tables (e.g., Table 3 looking at the interest rate) should be interpreted as the weighted average over the other treatments. None of the tables in the main paper or the appendix estimate the long model. We re-estimate the results in Table 3 (Column 1) and Table 8 (Column 1) including the interaction between the interest rate and the example maturity.

A.10 Education, HIV, and Early Fertility: Experimental Evidence from Kenya

Duflo et al. (2015a) analyze a field experiment with three interventions: education subsidies, HIV education, and a "critical think" intervention in which students are promoted to organized a debate and write an essay about condoms and HIV prevention. The first two treatments are implemented in a factorial design, and the authors include treatment dummies for each treatment as well as for the joint treatment. The third treatment is layered on top of schools that receive the HIV education, and while some Tables include the full treatment specification, the main tables do not. As the authors state: "For brevity, we ignore the randomized critical thinking (CT) intervention among H and SH schools in the main analysis (Tables 2, 3, and 4). We show the CT results in Table 5" We re-estimate Table 3: Column 4 and Table 4:Column 2 of the paper using the long model.[25]

A.11 Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment

Eriksson & Rooth (2014) study whether long-term unemployment spells matter for employers hiring decisions using a field experiment. The experimental design varies several applicant characteristics. Explicitly, "[t]he applicants were randomly assigned a number of attributes which typically are included in job applications and are expected to be important for the probability of being invited to a job interview. These attributes include contemporary and past spells of unemployment, work experience, education, gender, ethnicity, and some other characteristics." Each application was randomly assigned different characteristics using a factorial design. The following characteristics (and their possible values) were randomized: 1) Unemployment duration (takes value 0, 3, 6, or 9), 2) unemployed before employment (takes values 0 or 1), 3) unemployed between jobs

[24] We ignore this randomization since this is akin to a two-stage randomization designs, such as the one featured in Cohen & Dupas (2010), Karlan & Zinman (2009), or Ashraf et al. (2010).

[25] Since Critical Thinking took place 2 years after the other interventions, we focus on long-run outcomes.

(takes values 0 or 1), 4) work experience (takes values 1, 2, 3, 4 or 5), 5) number of employers (takes values 0 or 1), 6) ethnicity/gender (the applicant randomized to be native male, native female or ethnic minority male), 7) having more education than required (takes values 0 or 1), 8) work experience during the summer breaks (takes values 0 or 1), 9) visiting US high school (takes values 0 or 1), 10) Personality trait I - agency (takes values 0 or 1), 11) Personality trait II - communion (takes values 0 or 1), and 12) leisure activities (Randomized to have one of seven different leisure activities or none). As the authors explicitly state: "The typical approach in field experiments using the correspondence testing methodology is to vary only one characteristic in the applications, e.g., the ethnicity or gender of the applicant (cf. Riach and Rich 2002; Carlsson and Rooth 2007). However, in our experiment, we used a more general approach by randomly varying several characteristics. This allows us to measure the labor market return of different skills and attributes (cf. Bertrand and Mullainathan 2004; Rooth 2011)." The paper does not mention that the treatment effects in the main tables (e.g., Table 6) should be interpreted as the weighted average over the other treatments, nor does it estimate the full model in the paper or in the appendix. We re-estimate Table 6: Column 1 using the long model including all possible two way interactions, but there are higher order interactions implied by the factorial design.

A.12 Voting to Tell Others

DellaVigna et al. (2016) analyze the results from a field experiment designed to estimate a model of voting "because others will ask". To do this, they use a factorial design with four dimensions. First, households were randomized into five flyer treatments with equal weights, where the information received in a flyer varied across treatments. Then, they randomized the duration of the survey (5 minutes or 10 minutes). The third dimension, randomized how the surveyors described the survey to the respondent. The fourth dimension, randomized the incentives to a question regarding voting turnout. Figure A.5 — taken from the published version of the paper — shows the details of the randomization design. We replicate Table 1 (Columns 1 and 3) in the original paper including the interaction terms across treatments. Since the third and fourth randomization only take place after the respondent opens the door (which is the outcome we focus on) we focus on the first three dimensions. While the reduced form estimates in the paper do not estimate the long model, the paper also has a structural model that implicitly includes all the interactions in the estimation. However, the paper does not mention that the treatment effects in the main tables (e.g., Table 1) should be interpreted

as the weighted average over the other treatments.

Figure A.5: Factorial design in DellaVigna et al. (2016)

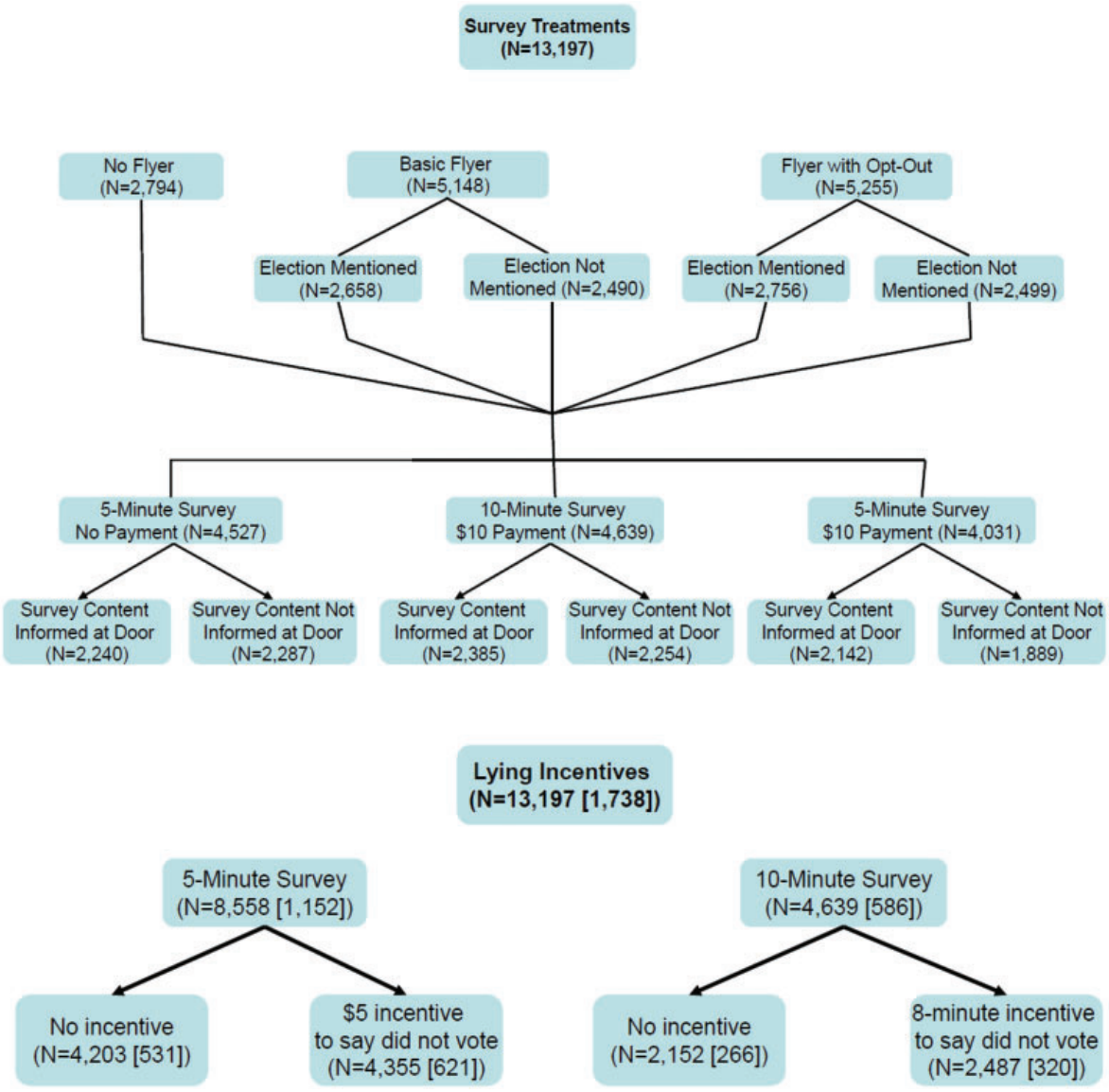


FIGURE 3

Experimental treatments

Note: Figure 3 presents the crossed experimental randomizations, with sample sizes in parentheses. On top are the five arms of the flyer treatment, crossed with whether respondents at the door are informed that the survey is about participation in the 2010 congressional election, crossed with survey duration and payment. At the bottom are the arms of the lying incentives, indicating both the initial sample size and [in square brackets] the sample size among individuals who responded to the survey. All arms are equally weighted and crossed.

Note: Figure 3 from DellaVigna et al. (2016).

A.13 What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods

Balafoutas et al. (2013) analyze a field experiment on taxi rides in Athens, Greece. The experiment is set up to measure fraud and to examine the influence of passengers' observable characteristics on fraud. The experiment vary the characteristics of passengers different taxi drivers got along two dimensions. First, passengers appear to be either local, nonlocal natives, or foreigner. Passengers in the roles of locals and non-local natives spoke in Greek, whereas passengers in the role of foreigners spoke in English. Passengers in the role of non-local natives and of foreigners asked the driver whether he knew the destination, adding as an explanation for asking that they were not familiar with the city. In addition, each passenger also appear to be either high- or low-income. Passengers intended to be perceived as having high income were dressed in a suit and carried a briefcase, whereas low-income passengers were dressed casually and carried a backpack. Figure A.6 — taken from the published version of the paper — shows the details of the randomization design. The paper does not mention that the treatment effects in the main tables (e.g., Table 5) should be interpreted as the weighted average over the other treatments. None of the tables in the main paper or the appendix estimate the long model. We re-estimate Table 5 (Columns 1-3) in the original paper including the interaction terms across treatments.

Figure A.6: Factorial design in Balafoutas et al. (2013)

TABLE 1
Treatments and locations in the experiment

[A] Treatments and number of observations

Passenger's information role	Passenger's income role		
	Low income	High income	Total
Local	58	58	116
Non-local native	58	58	116
Foreigner	58	58	116
Total	174	174	348

Note: Table 1 from Balafoutas et al. (2013).

A.14 The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya

Haushofer & Shapiro (2016) analyze a field experiment in which unconditional cash transfers are given to poor households. The experiment varies the unconditional cash transfers along three dimension: 1) whether the transfer is given to the primary female or the primary male in the household, 2) whether the transfers are given lump-sum or in monthly installments, and 3) the size of the transfer. The data is not available in the journal's website, but is available in the author's website.[26] Figure A.7 — taken from the published version of the paper — shows the details of the randomization design. The paper's main results (in Table 2) assume away spillovers and label the difference between the treatment and the spillover group as the treatment effect. The table shows the aggregate difference between all the treatment groups and the spillover group (Column 2), as well as the treatment difference across male vs female recipients (Column 3), monthly vs lump-sum transfers (Column 4), and large vs small transfers (Column 5). However, the results in Column 3-5 do not take into account the interactions between these treatments. The paper does not mention that the treatment effects in the main tables (e.g., Table 2) should be interpreted as the weighted average over the other treatments. None of the tables in the main paper or the appendix estimate the long model. Thus, we re-estimate all the estimates in Columns 3 to 5 of Table 2 including all the interactions between treatments.

[26]The data can be found at http://princeton.edu/haushofer

Figure A.7: Factorial design in Haushofer & Shapiro (2016)

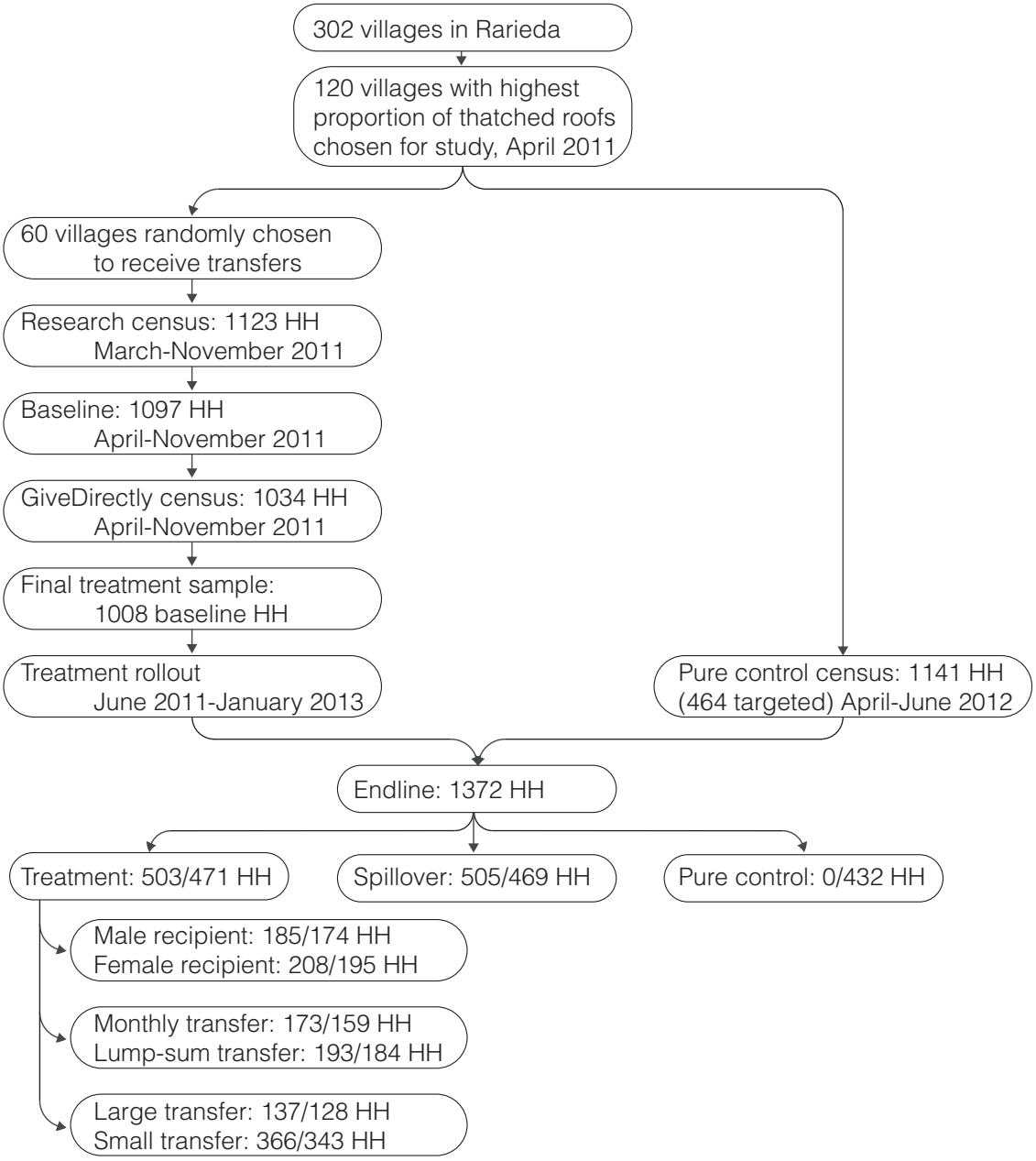


FIGURE I

Timeline of Study

Timeline and treatment arms. Numbers with slashes designate baseline/endline number of households in each treatment arm. Male versus female recipient was randomized only for households with cohabitating couples. Large transfers were administered by making additional transfers to households that had previously been assigned to treatment. The lump-sum versus monthly comparison is restricted to small transfer recipient households.

Note: Figure 1 from Haushofer & Shapiro (2016).

A.15 Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors

Khan et al. (2015) analyze an experiment in which tax collectors are paid for performance. This experiment features a 4x2 design. In the first dimension, units are assigned to either control, information only, or three different bonus schemes (+ information). In the second dimension, units are assigned to either control or performance pay for senior tax official. The results for the second randomization (i.e., performance pay for senior official) are no in the paper. In addition, the interactions are not included in the estimating equations. The data is not available in the journal's website, but is available in the author's website.[27]

The second treatment (incentives for senior officials) only took place during the second year of the experiment. The paper does not mention that the treatment effects in the main tables (e.g., Table 3) should be interpreted as the weighted average over the "senior officials treatment status". None of the tables in the main paper or the appendix estimate the long model. Thus, we re-estimator all the estimates in Columns 4 to 6 of Table 3 (Panel B) including all the interactions between treatments.[28]

A.16 Why the Referential Treatment? Evidence from Field Experiments on Referrals

Pallais & Sands (2016) analyzes three field experiments in an online labor market to study why referred workers are more likely than nonreferred workers to be hired. The same sample is randomized in three dimensions (the three experiments). The paper does not mention that the treatment effects in the main tables should be interpreted as the weighted average over the other treatments. None of the tables in the main paper or the appendix estimate the long model. The data is not available online.

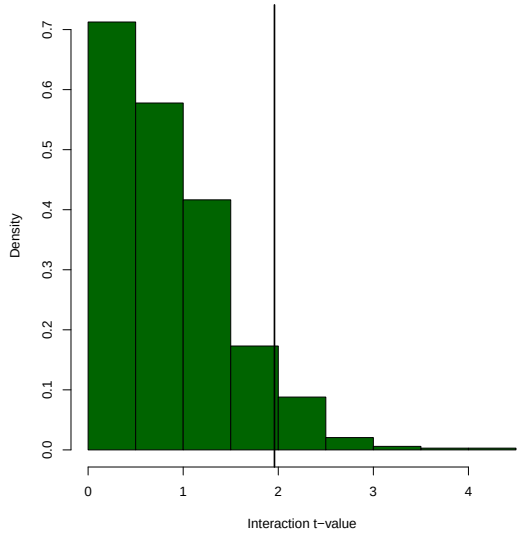
[27] The data can be found at https://economics.mit.edu/faculty/bolken/data

[28] The estimating equation used in the paper does not include a dummy variable for the information treatment, nor for the senior official treatment. We include both in our estimating equation *without* interactions.

A.17 Summary

A.17.1 All Papers

Figure A.8: Distribution of the t-value of interaction terms across studies - All

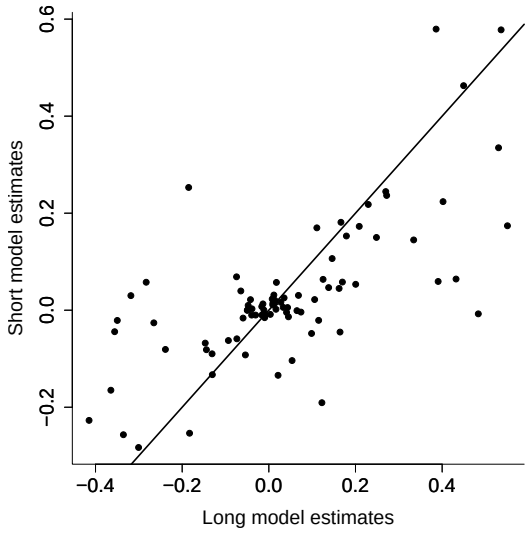


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.

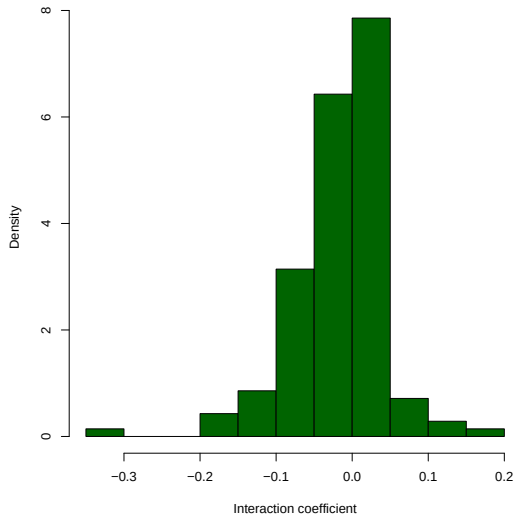
A.17.2 Ten Most Cited Papers

Figure A.9: Treatment estimates from the long and the short regression

(a) Main treatment estimates



(b) Interaction



Note: Both figures show treatment estimates from the ten most cited papers with factorial designs and publicly available data that do not include the interaction in the original study. Figure A.9a shows how the main treatment estimates change across the short and the long model across studies. The median main treatment estimate from the short model is 0.007σ, the median main treatment estimate from the long model is 0.014σ, 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 130.7%, and 28.2% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.9b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is -0.016σ, 4.9% of interactions are significant at the 10% level, 2.1% are significant at the 5% level, and 0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.55.

Table A.2: Significance of treatment estimates from the long and the short regression

Panel A: Significance at the 10% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	49	13	62
Significant	6	17	23
Total	55	30	85

Panel B: Significance at the 5% level

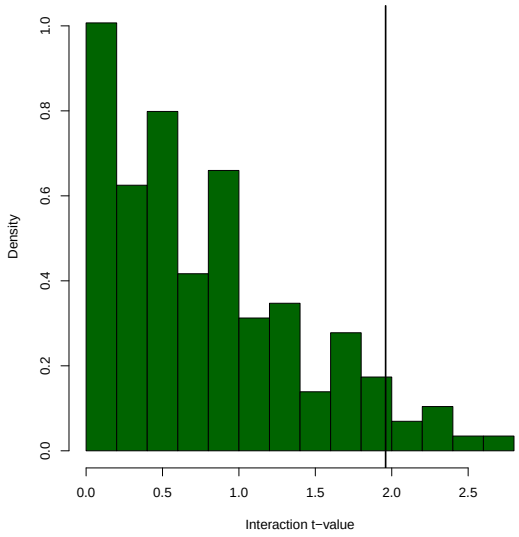
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	60	9	69
Significant	4	12	16
Total	64	21	85

Panel C: Significance at the 1% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	73	3	76
Significant	1	8	9
Total	74	11	85

This table shows the number of coefficients that are significant at a given level when estimating the long regression (columns) and the short regression (rows). This table only includes information from the ten most cited papers with factorial designs and publicly available data that do not include the interaction in the original study. Table 3 has data for all papers with factorial designs and publicly available data that do not include the interaction in the original study. Panel A uses a 10% significance level, Panel B uses 5%, and Panel C uses 1%.

Figure A.10: Distribution of the t-value of interaction terms across studies - ten most cited

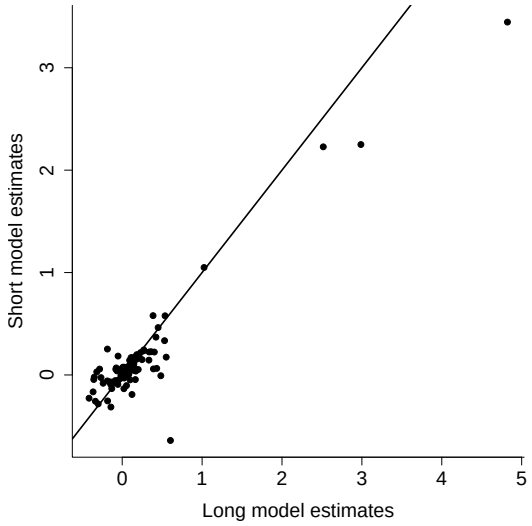


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.

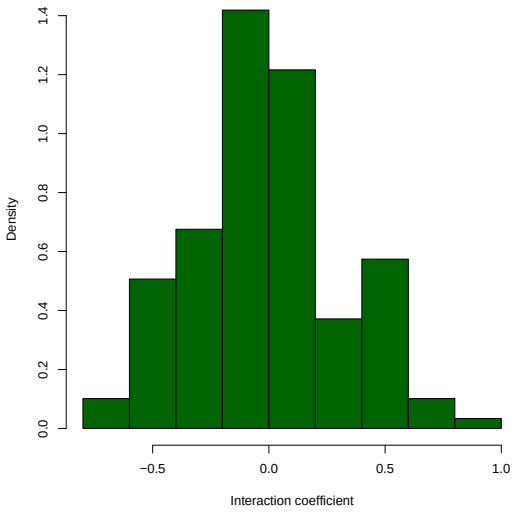
A.17.3 Papers with less than 20 Interactions

Figure A.11: Treatment estimates from the long and the short regression

(a) Main treatment estimates



(b) Interaction



Note: Both figures show treatment estimates from the papers with factorial designs and publicly available data that do not include the interaction in the original study and have less than 20 interactions. Figure A.11a shows how the main treatment estimates change across the short and the long model across studies. The median main treatment estimate from the short model is 0.051σ, the median main treatment estimate from the long model is 0.054σ, the average absolute difference between the treatment estimates of the short and the long model is 0.069σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 66%, and 22% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.11b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is -0.024σ, 5.9% of interactions are significant at the 10% level, 2.6% are significant at the 5% level, and 0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.54.

Table A.3: Significance of treatment estimates from the long and the short regression

Panel A: Significance at the 10% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	45	13	58
Significant	6	27	33
Total	51	40	91

Panel B: Significance at the 5% level

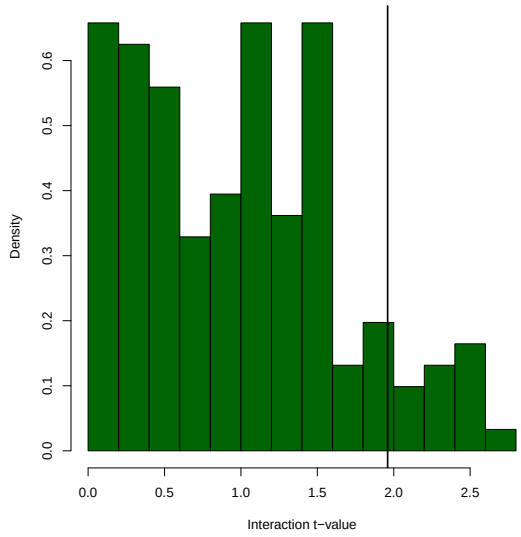
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	57	10	67
Significant	6	18	24
Total	63	28	91

Panel C: Significance at the 1% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	74	6	80
Significant	2	9	11
Total	76	15	91

This table shows the number of coefficients that are significant at a given level when estimating the long regression (columns) and the short regression (rows). This table only includes information from papers with factorial designs and publicly available data that do not include the interaction in the original study and have less than 20 interactions. Table 3 has data for all papers with factorial designs and publicly available data that do not include the interaction in the original study. Panel A uses a 10% significance level, Panel B uses 5%, and Panel C uses 1%.

Figure A.12: Distribution of the t-value of interaction terms across studies - ten most cited



Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.

B Effective Sample Size in Running Example

The difference in sample means for two groups (our estimate of the average treatment effect) is the OLS estimate of β in the regression

$$Y_i = \alpha + \beta T + \varepsilon \tag{21}$$

With independent and identically distributed data at the individual-level on Y, the variance of $\hat{\beta}$ is

$$\frac{\sigma^2}{P(1-P)N}$$

where σ^2 is the variance of Y, N is the total number of observations, and P is the proportion of individuals treated. If standard errors are clustered, the variance of the $\hat{\beta}$ is

$$\frac{\sigma^2}{P(1-P)J}\left(\rho + \frac{1-\rho}{n}\right)$$

where n is the number of observations per cluster, J is the number of clusters, and ρ is the intra-cluster correlation. We are interesting in knowing how many individual-level (independent) observations N, are equivalent in terms of power (i.e., yield the same variance for β) to J clusters and n observations per cluster.

$$\begin{aligned}
\frac{\sigma^2}{P(1-P)J}\left(\rho + \frac{1-\rho}{n}\right) &= \frac{\sigma^2}{P(1-P)N} \\
\frac{1}{J}\left(\rho + \frac{1-\rho}{n}\right) &= \frac{1}{N} \\
J\left(\frac{n}{(n-1)\rho + 1}\right) &= N
\end{aligned}$$

Therefore a cluster randomized experiment with $\rho = 0.3$, $J = 350$, and $n = 30$ is equivalent to an individual level experiment with $N = 1,082$. We round this number to $1,100$.

C Derivations

C.1 Derivation of the expressions for β_1^s and β_2^s

We consider Equation (10). The derivation of Equation (11) is similar and thus omitted. One can write the observed outcome as

$$Y = Y_{0,0} + (Y_{1,0} - Y_{0,0})T_1(1 - T_2) + (Y_{0,1} - Y_{0,0})(1 - T_1)T_2 + (Y_{1,1} - Y_{0,0})T_1T_2 \tag{22}$$

Next, consider

$$\begin{aligned}
E\left(Y \mid T_1 = 1\right) &= E\left(Y_{0,0} + (Y_{1,0} - Y_{0,0})(1 - T_2) + (Y_{1,1} - Y_{0,0})T_2 \mid T_1 = 1\right) \\
&= E\left(Y_{0,0} + (Y_{1,0} - Y_{0,0})(1 - T_2) + (Y_{1,1} - Y_{0,0})T_2\right) \\
&= E\left(Y_{1,0} - Y_{0,0}\right) + E\left(Y_{0,0} - (Y_{1,0} - Y_{0,0})T_2 + (Y_{1,1} - Y_{0,0})T_2\right)
\end{aligned}$$

where the first equality follows by (22) and second equality follows by the independence Assumption 1. Similarly,

$$\begin{aligned}
E\left(Y \mid T_0 = 0\right) &= E\left(Y_{0,0} + (Y_{0,1} - Y_{0,0})T_2 \mid T_1 = 0\right) \\
&= E\left(Y_{0,0} + (Y_{0,1} - Y_{0,0})T_2\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
E(Y \mid T_1 = 1) - E(Y \mid T_1 = 0) &= E(Y_{1,0} - Y_{0,0}) + E(Y_{0,0} - (Y_{1,0} - Y_{0,0})T_2 + (Y_{1,1} - Y_{0,0})T_2) \\
&\quad - E(Y_{0,0} + (Y_{0,1} - Y_{0,0})T_2) \\
&= E(Y_{1,0} - Y_{0,0}) \\
&\quad + E((Y_{1,1} - Y_{0,0})T_2 - (Y_{1,0} - Y_{0,0})T_2 - (Y_{0,1} - Y_{0,0})T_2) \\
&= E(Y_{1,0} - Y_{0,0}) + E((Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0})\, T_2) \\
&= E(Y_{1,0} - Y_{0,0}) + E(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0} \mid T_2 = 1)\, P(T_2 = 1) \\
&= E(Y_{1,0} - Y_{0,0}) + E(Y_{1,1} - Y_{1,0} - Y_{0,1} + Y_{0,0})\, P(T_2 = 1),
\end{aligned}$$

where the fourth equality follows by the law of iterated expectations and the fifth equation follows by Assumption 1.

Next, we derive Equation (12). Consider

$$\begin{aligned}
E(Y \mid T_1 = 1) - E(Y \mid T_1 = 0) &= E(Y_{1,1})\, P(T_2 = 1) + E(Y_{1,0})\, P(T_2 = 0) \\
&\quad - [E(Y_{0,1})\, P(T_2 = 1) + E(Y_{0,0})\, P(T_2 = 0)] \\
&= E(Y_{1,1} - Y_{0,1})\, P(T_2 = 1) + E(Y_{1,0} - Y_{0,0})\, P(T_2 = 0),
\end{aligned}$$

where the first equality follows from the law of iterated expectations and Assumption 1. The result now follows from from Equations (5)–(7). The derivation of Equation (13) is similar and thus omitted.

C.2 Derivation of the expressions for β_1^s and β_2^s when the interaction cell is empty

Here we derive the expressions for β_1^s and β_2^s if we leave the interaction cell empty, i.e., $P(T_1 = 1, T_2 = 1) = 0$. To simplify the exposition, we define $p_1 := P(T_1 = 1)$, $p_2 := P(T_2 = 1)$ and $p_{12} := P(T_1 = 1, T_2 = 1)$. The population regression coefficients $\beta^s = (\beta_0^s, \beta_1^s, \beta_2^s)'$ is given by

$$\beta^s = E\left(XX'\right)^{-1} E(XY),$$

where $X = (1, T_1, T_2)'$. Multiplying out yields the following expressions for β_1^s and β_2^s:

$$\begin{aligned}
\beta_1^s &= \frac{(p_2 p_{12} - p_1 p_2)E(Y) + p_1(p_2 - p_2^2)E(Y \mid T_1 = 1) + p_2(p_1 p_2 - p_{12})E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2 + 2p_1 p_2 p_{12} - p_{12}^2}, \\
\beta_2^s &= \frac{(p_1 p_{12} - p_1 p_2)E(Y) + p_1(p_1 p_2 - p_{12})E(Y \mid T_1 = 1) + p_2(p_1 - p_1^2)E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2 + 2p_1 p_2 p_{12} - p_{12}^2},
\end{aligned}$$

In what follows, we focus on β_1^s; the derivation for β_2^s is similar. Using $p_{12} = 0$, obtain

$$\beta_s^1 = \frac{-p_1 p_2 E(Y) + p_1 p_2(1 - p_2)E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \tag{23}$$

Because $p_{12} = 0$, we have that

$$E(Y) = E(Y \mid T_1 = 1, T_2 = 0)p_1 + E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2) + E(Y \mid T_1 = 0, T_2 = 1)p_2. \tag{24}$$

Combining (23) and (24) and simplifying the yields the desired result:

$$\begin{aligned}
\beta_1^s &= \frac{-p_1 p_2(E(Y \mid T_1 = 1, T_2 = 0)p_1 + E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \\
&\quad - \frac{E(Y \mid T_1 = 0, T_2 = 1)p_2) + p_1 p_2(1 - p_2)E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \\
&= \frac{p_1 p_2 E(Y \mid T_1 = 1, T_2 = 0)(1 - p_1 - p_2) - p_1 p_2 E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \\
&= \frac{E(Y \mid T_1 = 1, T_2 = 0)(1 - p_1 - p_2) - E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2)}{(1 - p_1 - p_2)} \\
&= E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0)
\end{aligned}$$

D Variance reductions and power gains based on the short model

D.1 Formal analysis

The following lemma computes and compares the finite sample power of a two-sided t-test for the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ based on the short and the long regression model. Recall from Section 2.3 that $Var(\hat{\beta}_1) > Var(\hat{\beta}_1^s)$. Here we show that the lower variance of $\hat{\beta}_1^s$ leads to higher power when $\beta_{12} = 0$.

Define $\boldsymbol{T}_1 \equiv (T_{11}, \ldots, T_{1N})'$ and $\boldsymbol{T}_2 \equiv (T_{21}, \ldots, T_{2N})'$, let $\hat{t}^s = \hat{\beta}_1^s / SE\left(\hat{\beta}_1^s\right)$ and $\hat{t} =$

$\hat{\beta}_1 / SE\left(\hat{\beta}_1\right)$, let P_{β_1} denote probabilities under the assumption that β_1 is the true coefficient and let $c_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, where Φ^{-1} is the quantile function of the standard normal distribution and $\alpha \in (0, 0.5)$ is the significance level.

Lemma 1. *Suppose that* $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_N) \mid \boldsymbol{T}_1, \boldsymbol{T}_2 \sim N(0, \sigma^2 I_N)$ *and* $\beta_{12} = 0$. *Then:*

(i) The finite sample power of the t-tests based on the short and the long model are given as

$$P_{\beta_1}\left(|\hat{t}| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) = \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1\right)} - c_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1\right)} + c_{1-\alpha/2}\right),$$

and

$$P_{\beta_1}\left(|\hat{t}^s| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) = \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1^s\right)} - c_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1^s\right)} + c_{1-\alpha/2}\right).$$

(ii) The t-test based on the short model is more powerful than the t-test based on the long model:

$$P_{\beta_1}\left(|\hat{t}^s| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) \geq P_{\beta_1}\left(|\hat{t}| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right).$$

Proof. **Part (i):** Under the assumptions in the statement of the lemma,

$$\frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} \mid \boldsymbol{T}_1, \boldsymbol{T}_2 \sim N(0,1)$$

It follows that, for $z \in \mathbb{R}$,

$$\begin{aligned} P_{\beta_1}\left(\hat{t} > z \,\middle|\, \boldsymbol{T}_1, \boldsymbol{T}_2\right) &= P_{\beta_1}\left(\frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} > z \,\middle|\, \boldsymbol{T}_1, \boldsymbol{T}_2\right) \\ &= P_{\beta_1}\left(\frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} > z - \frac{\beta_1}{SE\left(\hat{\beta}_1\right)} \,\middle|\, \boldsymbol{T}_1, \boldsymbol{T}_2\right) \\ &= \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1\right)} - z\right). \end{aligned}$$

Thus, the power of a two-sided test is

$$\begin{aligned} P_{\beta_1}\left(|\hat{t}| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) &= P_{\beta_1}\left(\hat{t} > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) + P_{\beta_1}\left(\hat{t} < -c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) \\ &= \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1\right)} - c_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1\right)} + c_{1-\alpha/2}\right) \end{aligned}$$

Similarly, one can show that

$$P_{\beta_1}\left(|\hat{t}^s| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) = \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1^s\right)} - c_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\beta_1}{SE\left(\hat{\beta}_1^s\right)} + c_{1-\alpha/2}\right)$$

Part (ii): Using the same arguments as in Part (i), it follows that the power of a t-test based on an estimator $\tilde{\beta}_1$ which satisfies

$$\tilde{t} := \frac{\tilde{\beta}_1 - \beta_1}{SE(\tilde{\beta})} \mid \boldsymbol{T}_1, \boldsymbol{T}_2 \sim N(0,1)$$

is given by

$$P_{\beta_1}\left(|\tilde{t}| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right) = \Phi\left(\frac{\beta_1}{SE(\tilde{\beta})} - c_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\beta_1}{SE(\tilde{\beta})} + c_{1-\alpha/2}\right).$$

Consider[29]

$$\begin{aligned}\frac{\partial P_{\beta_1}\left(|\tilde{t}| > c_{1-\alpha/2} \mid \boldsymbol{T}_1, \boldsymbol{T}_2\right)}{\partial SE(\tilde{\beta})} &= \phi\left(\frac{\beta_1}{SE(\tilde{\beta})} - c_{1-\alpha/2}\right)\frac{-\beta_1}{SE(\tilde{\beta})^2} - \phi\left(\frac{\beta_1}{SE(\tilde{\beta})} + c_{1-\alpha/2}\right)\frac{-\beta_1}{SE(\tilde{\beta})^2} \\ &= \frac{\beta_1}{SE(\tilde{\beta})^2}\left[\phi\left(\frac{\beta_1}{SE(\tilde{\beta})} + c_{1-\alpha/2}\right) - \phi\left(\frac{\beta_1}{SE(\tilde{\beta})} - c_{1-\alpha/2}\right)\right] \leq 0,\end{aligned}$$

which follows from the shape of the normal distribution. The desired result now follows by noting that $SE(\hat{\beta}_1^s) < SE(\hat{\beta}_1)$. □

D.2 Power gains and the size of the interaction cell

Here we discuss how the power gains of the t-test based on the short model are related to the size of the interaction cell. Recall from Section 2.3 that, in a 2×2 factorial design, the variance of the estimate of β_1 is given by

$$Var\left(\hat{\beta}_1\right) \equiv \sigma^2\frac{N_1 + N_2}{N_1 N_2} \quad \text{and} \quad Var\left(\hat{\beta}_1^s\right) \equiv \sigma^2\frac{N_1N_3 + N_1N_4 + N_2N_3 + N_2N_4}{N_1N_2N_3 + N_1N_2N_4 + N_1N_3N_4 + N_2N_3N_4}.$$

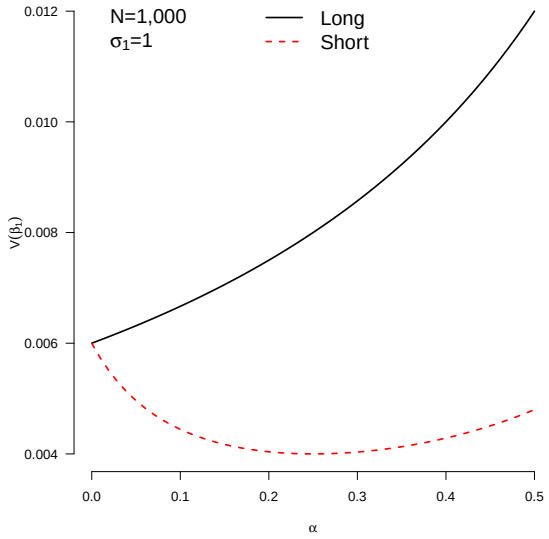
To showcase the difference in power of the t-test (which is inversely related to the variance) we simplify the problem by assuming that $N_1 = N_2 = N_3$, and hence that the researcher simply has to determine the relative size of N_4. Let α be such that $N_4 = \alpha N$. Thus, $N_1 = N_2 = N_3 = \frac{1}{3}(1-\alpha)N$. Then:

[29] See, for example, Lemma 2 in Carneiro et al. (2017) for a similar argument.

$$Var\left(\hat{\beta}_1\right) \equiv \sigma^2 \frac{6}{(1-\alpha)N} \quad \text{and} \quad Var\left(\hat{\beta}_1^s\right) \equiv \sigma^2 \frac{6(1+2\alpha)}{(1-\alpha)N(1+8\alpha)}.$$

Figure D.1 shows how the variance changes for different values of α. The more sample we allocate to the interaction cell, the higher the variance of $\hat{\beta}_1$ (i.e., the lower the power) of the long model. However, for the short model the relationship is non-monotonic. The highest power (lowest variance) is achieved when the sample size is allocated equally across cells (i.e., $\alpha = 0.25$). Intuitively, given that we ignore the fact that some individuals get both treatments, at this point the size of the treatment and the control group for T_1 is the same.

Figure D.1: $Var\left(\hat{\beta}_1\right)$ and $Var\left(\hat{\beta}_1^s\right)$ as the interaction cell becomes larger (N=1,000)



E Additional figures and tables

Table E.1: Articles publised in top-5 journals between 2006 and 2017

	AER	ECMA	JPE	QJE	ReStud	Total
Other	1218	678	367	445	563	3271
Field experiment	43	9	14	45	13	124
Lab experiment	61	16	5	10	18	110
Total	1322	703	386	500	594	3505

Table E.2: Significance of treatment estimates from the short regression and using the method outline in Section 5.2

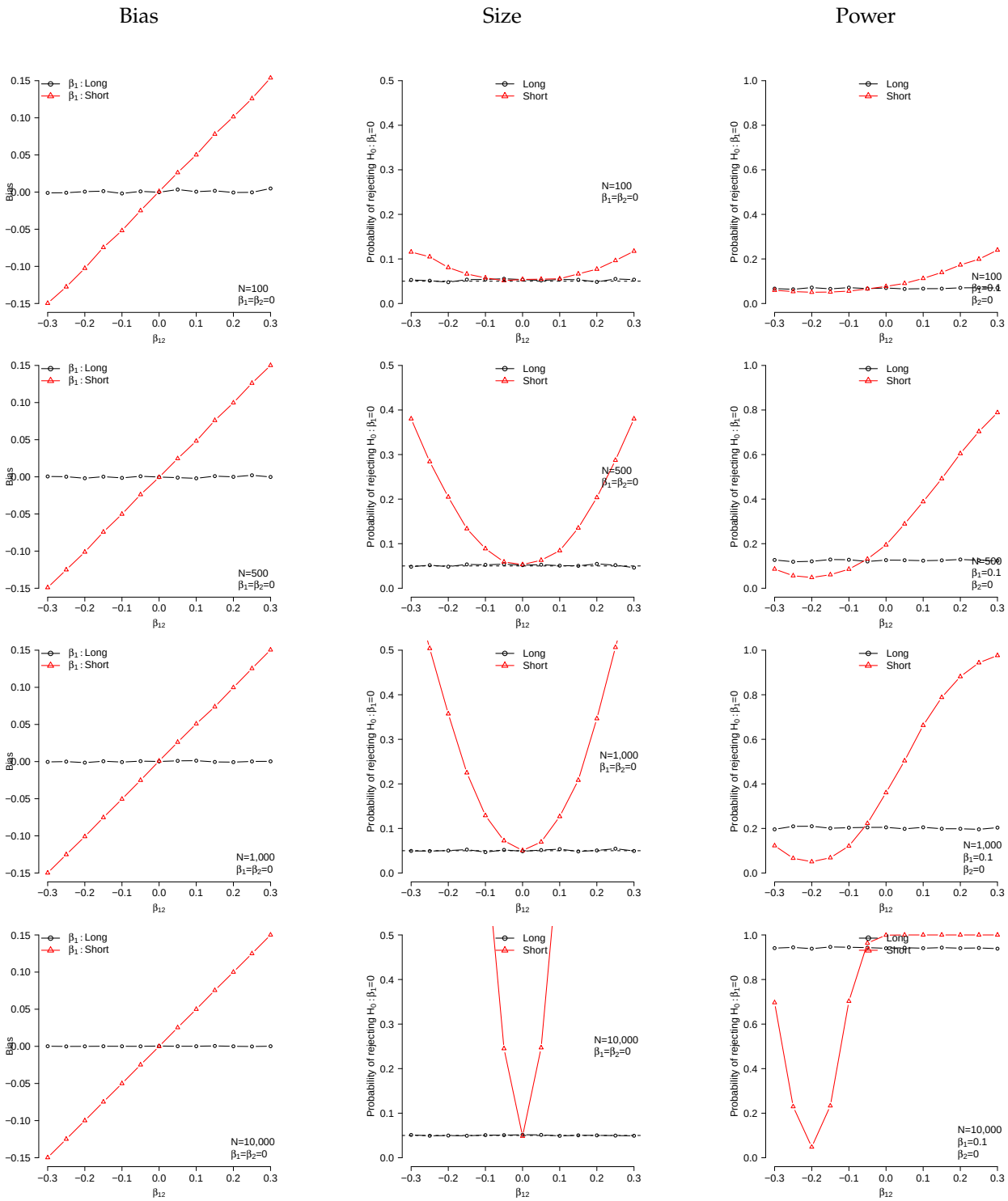
Panel A: Significance at the 5% level

	Without interaction		
Prior Knowledge	Not significant	Significant	Total
Not significant	62	24	86
Significant	0	7	7
Total	62	31	93

This table shows the number of coefficients that are significant at a given level using the method outline in Section 5.2 (columns) and the short regression (rows). This table only includes information from the ten most cited papers with factorial designs and publicly available data that do not include the interaction in the original study.

E.1 Ignoring the interaction

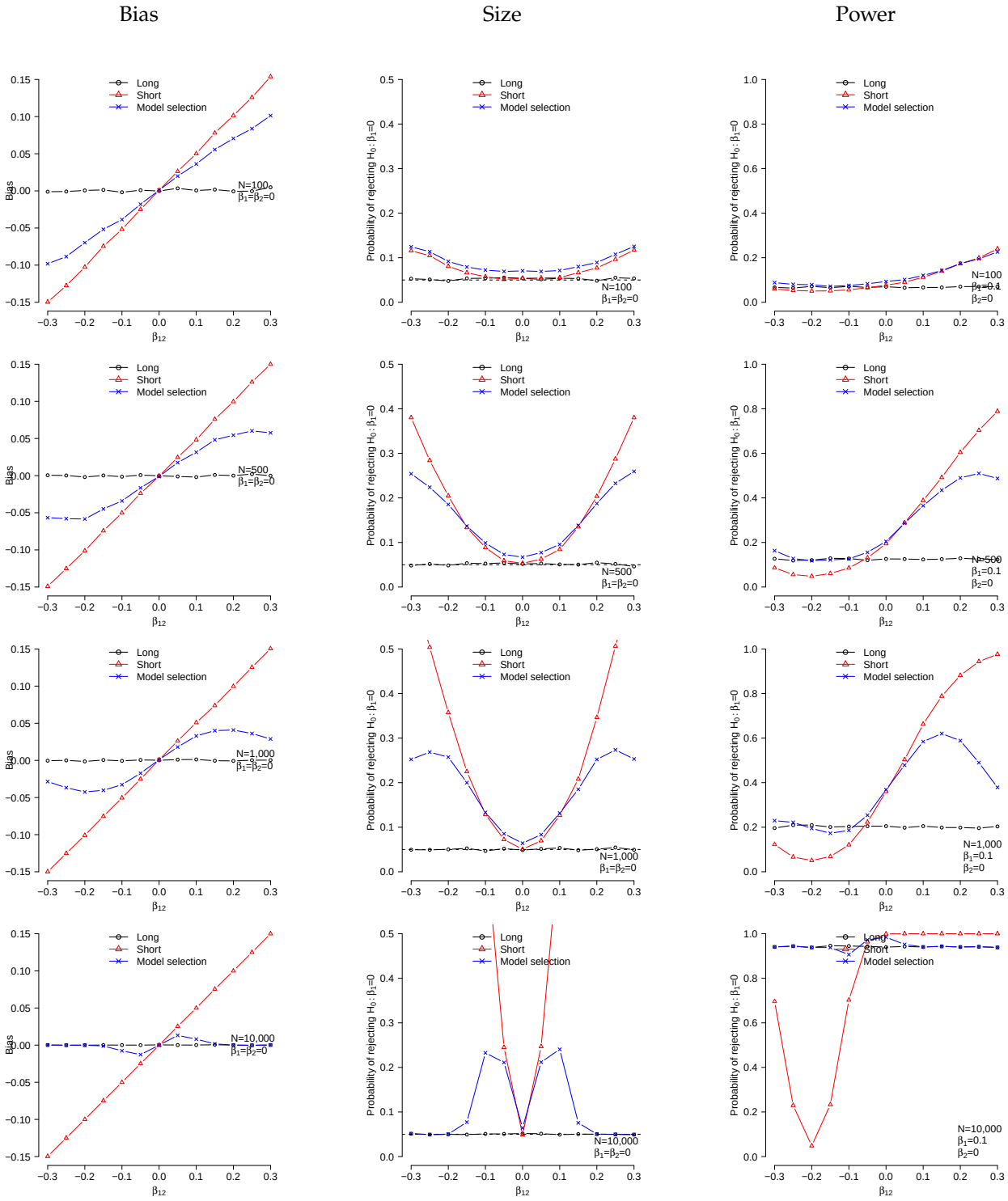
Figure E.1: Long and short model: Bias, size, and power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$.

E.2 Pre-testing

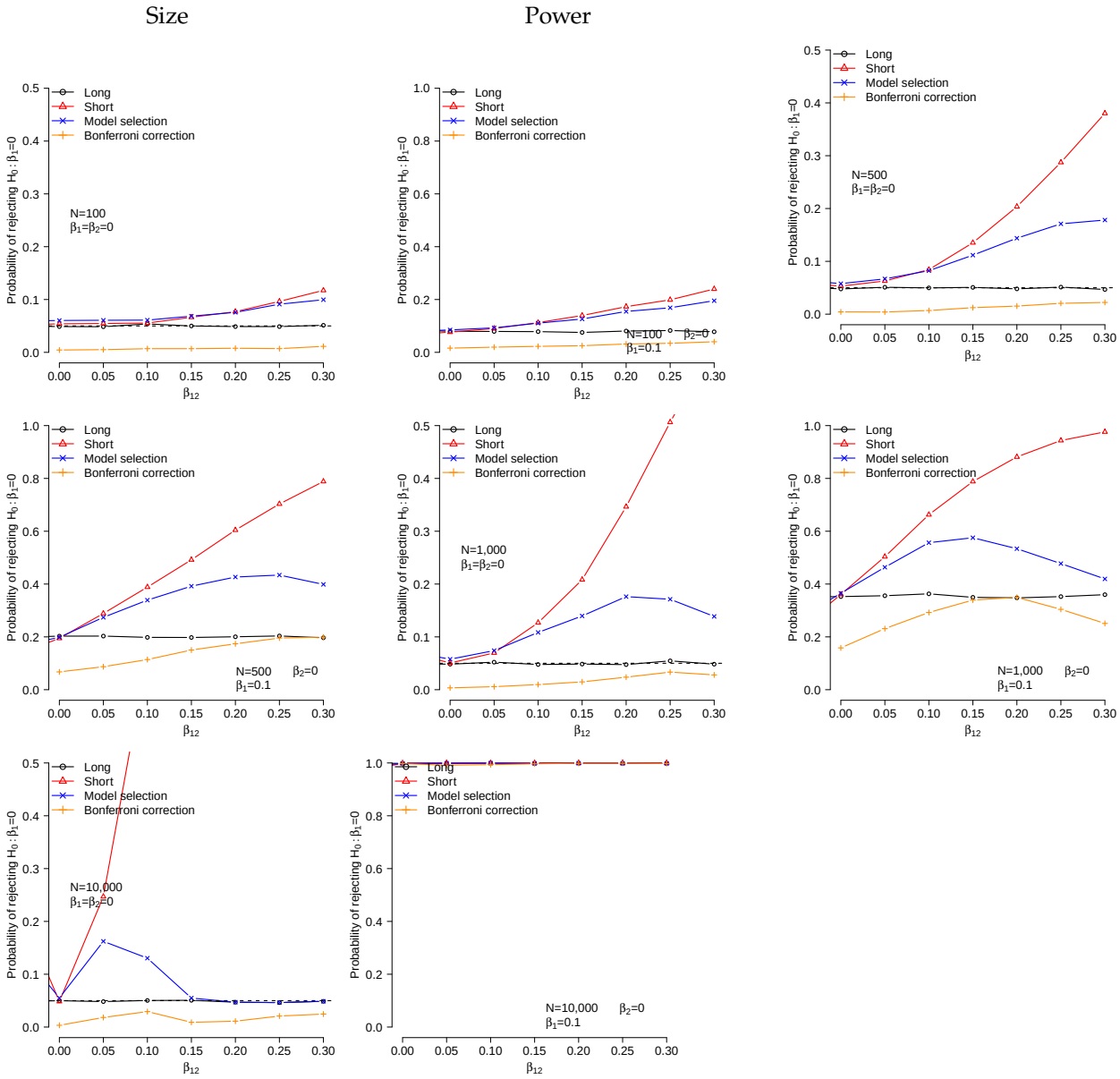
Figure E.2: Model selection: Bias, size, and power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures 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.

E.3 McCloskey (2017)'s Bonferroni-style Correction

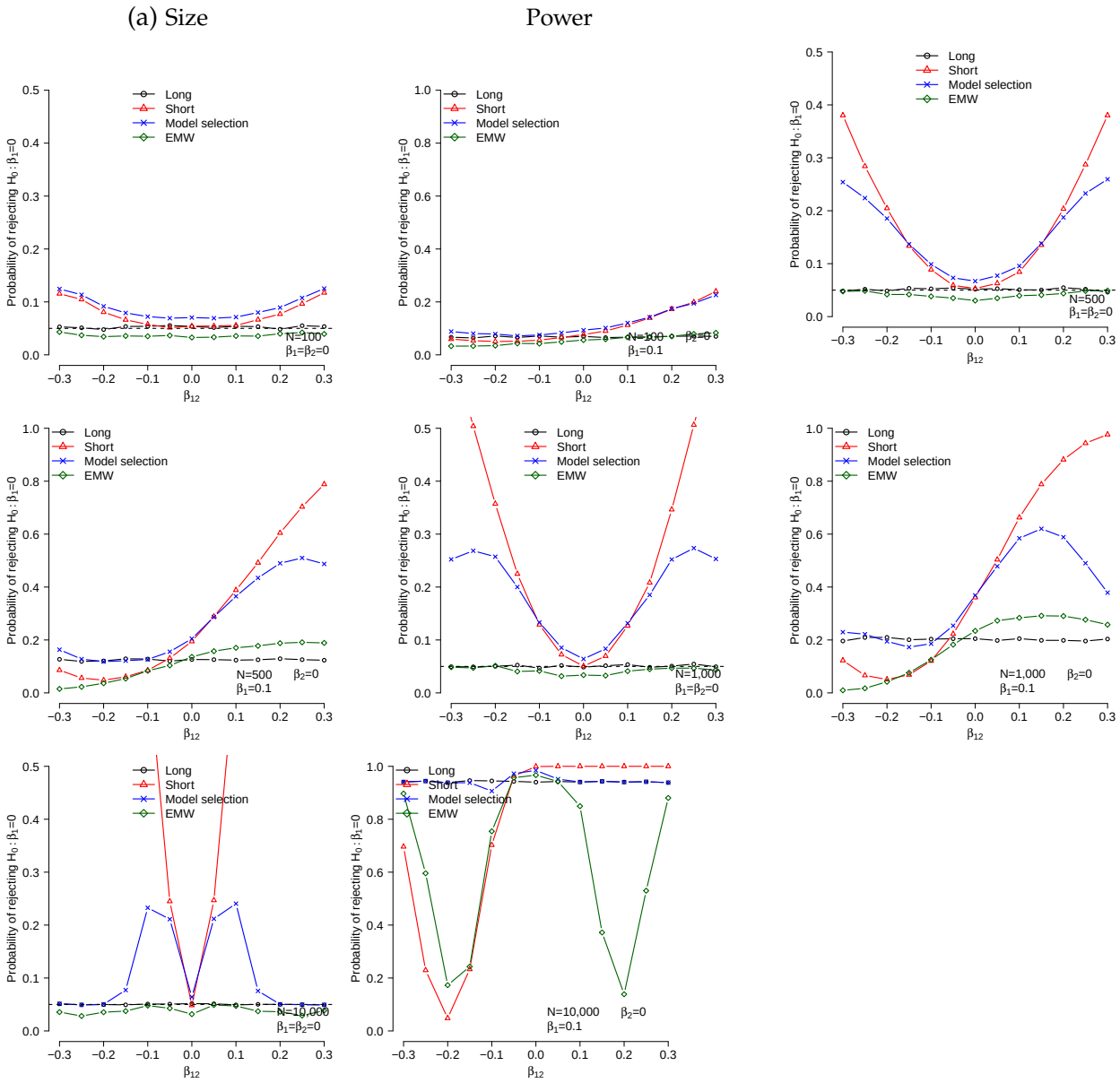
Figure E.3: McCloskey (2017)'s Bonferroni-style correction: Size and power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures 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.*

E.4 Elliott et al. (2015)'s nearly optimal test

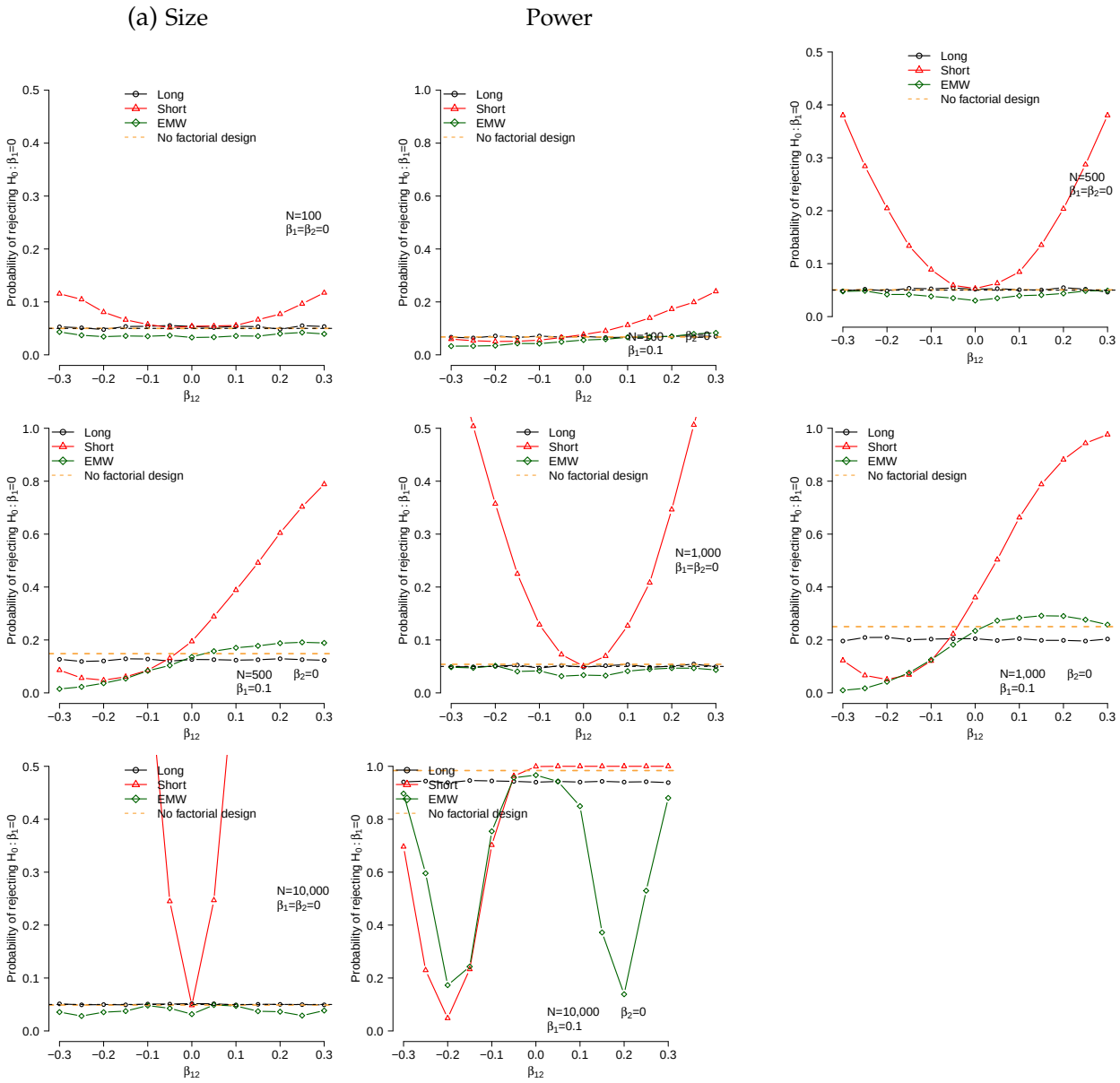
Figure E.4: Elliott et al. (2015)'s nearly optimal test: size and power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures 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.*

E.5 Leaving the interaction cell empty

Figure E.5: No factorial design: size and power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. *When there is no factorial design, we assume the* N_4 *individuals in the interaction cell are equally distributed to the other cells.*