NBER WORKING PAPER SERIES

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

Karthik Muralidharan
Mauricio Romero
Kaspar Wüthrich

Working Paper 26562
http://www.nber.org/papers/w26562

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2019

We are grateful to Isaiah Andrews, Tim Armstrong, Arun Chandrasekhar, Clement de Chaisemartin, Gordon Dahl, Stefano DellaVigna, Esther Duflo, Graham Elliott, Markus Goldstein, Macartan Humphreys, Hiroaki Kaido, Michal Kolesar, Soonwoo Kwon, Adam McCloskey, Craig McIntosh, Paul Niehaus, Rachael Meager, Ben Olken, Gautam Rao, Andres Santos, Jesse Shapiro, 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments
Karthik Muralidharan, Mauricio Romero, and Kaspar Wüthrich
NBER Working Paper No. 26562
December 2019
JEL No. C12,C18,C90,C93

ABSTRACT

Factorial designs are widely used for studying multiple treatments in one experiment. While "long" model t-tests provide valid inferences, t-tests using the "short" model (ignoring interactions) yield higher power if interactions are zero, but incorrect inferences otherwise. Of 27 factorial experiments published in top-5 journals (2007--2017), 19 use the short model. After including all interactions, over half their results lose significance. Modest local power improvements over the long model are possible, but with lower power for most values of the interaction. If interactions are not of interest, leaving the interaction cells empty yields valid inferences and global power improvements.

Karthik Muralidharan
Department of Economics, 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
kamurali@ucsd.edu

Kaspar Wüthrich
University of California at San Diego
kwuthrich@ucsd.edu

Mauricio Romero
Centro de Investigacion Economica
ITAM
Mexico
mtromero@itam.mx

An online appendix is available at http://www.nber.org/data-appendix/w26555

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 in the same experiment. 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 results from the short model, with the implicit assumption that the interaction is zero.

The power gains based on the short model, however, come at the cost of an increased likelihood of incorrect inference relative to a "business as usual" counterfactual if the interaction effect is not zero. 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, 19 do not include all interaction terms in the main specifications. We re-analyzed the data from these papers by also including the interaction terms.[1] Doing so has non-trivial implications for inference regarding the main treatment effects. The median absolute change in the point estimates of the main treatment effects is 96%, about 26% of estimates change sign, and 53% (29 out of 55) of estimates reported to be significant at the 5% level are no longer so after including interactions. Even if we re-analyze only "policy" experiments and exclude "conceptual" experiments (which have more treatments and interactions on average), the rate of false rejection is still 32%.[2]

[1] The full list of 27 papers is in Table A.1. We re-analyzed 15 out of the 19 that do not include all interactions in the main specification. The other four papers did not have publicly-accessible data.

[2] We define a policy experiment as one which studies a program or intervention that could be scaled up, and a conceptual experiment as one which aims to test for the existence of facts or concepts such as discrimination (studied, for instance, by resume audit experiments).

In practice, researchers often try to address the issue of interactions by first estimating the long model and testing if the interaction is significant, and then focusing on the short model if they do not reject that the interaction is zero. However, the distributions 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.[3] Thus, this two-step procedure will almost always fail to reject that the interaction term is zero, even when it is different from zero. As a result, the two-step procedure will typically not control size, and often lead to incorrect inferences regarding treatment effects against a business-as-usual counterfactual.

Textbook treatments of factorial designs (Cochran & Cox, 1957; Gerber & Green, 2012) and guides to practice (Kremer, 2003; Duflo et al., 2007) are careful to clarify that treatment effects using the short model should be interpreted as either (a) being conditional on the distribution of the other treatment arms in the experiment, or (b) as a composite treatment effect that includes a weighted-average of the interactions with other treatments. However, as we argue, this weighted average is a somewhat arbitrary construct, and typically neither of primary academic interest nor policy-relevant. Consistent with this view, *none* of the 19 experimental papers in our reanalysis that ignore the interactions motivate their experiment as being about estimating this composite weighted-average treatment effect, or caveat that the presented treatment effects should be interpreted as being against a counterfactual that also has the other treatments in the same experiment.

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

Second, interactions are quantitatively important. We find that the median absolute value of interactions relative to the main treatment effects is 0.37.[4] The view that interactions are second-order may have been influenced in part by the lack of evidence

[3]For example, Gelman (2018) shows that one 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 the main effects.

[4]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. This results in a non-trivial rate of false rejection of the null hypothesis of zero effects of the main treatment against a business-as-usual counterfactual in *any given study*.

of significant interactions in most experiments to date. However, this is at least partly because 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 (Duflo et al., 2015a; Mbiti et al., 2019) 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.

Third, factorial designs may make sense if the goal is not hypothesis testing but to minimize mean squared error (MSE) criteria (or other loss functions), where the researcher is willing to accept some bias for lower variance (e.g., Blair et al., 2019). However, policy experiments are typically used based on whether the intervention had a "significant" effect. This is both because of publication bias towards significant findings (e.g., I. Andrews & Kasy, 2018; Christensen & Miguel, 2018; Franco et al., 2014), and because meta-analyses and evidence 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 (which we document in this paper), is likely to have non-trivial implications for how evidence is published, summarized, and translated into policy.

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 that improve power relative to the long model *while maintaining size control* for relevant values of the interactions. Before summarizing these econometric possibilities, we note that the two-sided t-test based on the long model is the uniformly most powerful unbiased test (e.g., van der Vaart, 1998). This classical result implies that any procedure that is more powerful than the t-test for some values of the interactions, must underperform somewhere else. Moreover, even in the best case, the scope for power improvements is limited if one insists on size control for all values of the interactions.[5] Keeping these constraints in mind, we explore four possible econometric approaches.

The first approach, based on Elliott et al. (2015), is a nearly optimal test that targets power towards an a priori likely value of the interaction (such as a value of zero), while

[5]For the corresponding one-sided testing problem, the one-sided t-test is uniformly most powerful. Thus, the best one can hope for is to improve power from the two-sided to the one-sided test. This power improvement is never larger than 12.5 percentage points at the 5% level.

controlling size for all values of the interaction. We find that this approach comes close to achieving the maximal possible (modest) power gains near the likely values of the interaction, while exhibiting lower power farther away from this value. The nearly optimal test can be useful in 2×2 factorial designs with strong prior knowledge about the interactions, but becomes computationally prohibitive in more complicated factorial designs.[6] Our second approach, based on Armstrong et al. (2019), is to construct confidence intervals for the main effects under prior knowledge on the magnitude of the interactions. When the prior knowledge is correct, this approach controls size and yields power gains relative to the t-test based on the long model. However, it suffers 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. In the appendix, we explore two additional econometric approaches based on work by Imbens & Manski (2004), Stoye (2009) and McCloskey (2017). Our simulations show that these approaches are unlikely to yield meaningful power improvements relative to the first two approaches and the long model.

Based on the analysis above, we recommend that all completed factorial experiments report results from and use the t-test based on the long regression model. It is easy to compute even in complicated factorial designs and has appealing optimality properties. 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 focus on results from the short model, they should clearly indicate that treatment effects should be interpreted as a composite treatment effect that includes a weighted-average of interactions with other treatments (and commit to the estimand of interest in a pre-analysis plan). This will ensure transparency in the interpretation of the main results and enable readers to assess the extent to which the other treatments may or may not be typical background factors that can be ignored. Finally, we note that even in settings where the coefficients in the short model are of interest, they can always be constructed based on the coefficients in the long model, while the converse is not true.

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 the econometric methods discussed above for most of the

[6]Since most of the experiments that we re-analyze have more complex factorial designs, this procedure is of limited use for these experiments. Our code to implement this procedure for 2×2 factorial designs is available at https://mtromero.shinyapps.io/elliott/.

relevant values of the interaction effect. 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.

The recommendations above are most relevant for the design and analysis of policy experiments, where a business-as-usual counterfactual is important. Factorial designs, and analyses of the short model may be fine in conceptual experiments, such as resume audit studies, where many (or all) the characteristics that are randomized (such as age, education, race, and gender) do exist in the population. In these cases, a weighted average effect (as estimated by the short model) may be a reasonable target parameter subject to researchers indicating clearly how the resulting effect should be interpreted. However, even in these cases, there may be value in having the treatment share of various characteristics being studied be the same as their population proportion. Doing so will make the short-model coefficient more likely to approximate the population relevant parameter of interest.

Our most important 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 (2018), who shows that randomization tests result in 13% to 22% fewer significant 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.

2 Theoretical analysis 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; we do so in 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 Potential outcomes and treatment effects

We formalize the problem using the potential outcomes framework of Rubin (1974). Our goal is to identify and estimate the causal effect of the two treatments, T_1 and T_2, on an outcome of interest, Y. 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 (ATEs):

$$\begin{aligned}
E\left(Y_{1,0} - Y_{0,0}\right): &\quad \text{ATE of } T_1 \text{ relative to a counterfactual where } T_2 = 0 \\
E\left(Y_{0,1} - Y_{0,0}\right): &\quad \text{ATE of } T_2 \text{ relative to a counterfactual where } T_1 = 0 \\
E\left(Y_{1,1} - Y_{0,1}\right): &\quad \text{ATE of } T_1 \text{ relative to a counterfactual where } T_2 = 1 \\
E\left(Y_{1,1} - Y_{1,0}\right): &\quad \text{ATE of } T_2 \text{ relative to a counterfactual where } T_1 = 1 \\
E\left(Y_{1,1} - Y_{0,0}\right): &\quad \text{ATE of } T_1 \text{ and } T_2 \text{ combined}
\end{aligned}$$

We refer to $E\left(Y_{1,0} - Y_{0,0}\right)$ and $E\left(Y_{0,1} - Y_{0,0}\right)$ as the *main treatment effects* of T_1 and T_2 relative to a business-as-usual counterfactual where no one is affected by the treatments analyzed in the experiment. The interaction effect — the difference between the effect of jointly providing both treatments and the sum of the main effects — is

$$E\left(Y_{1,1}-Y_{0,0}\right)-\left[E\left(Y_{1,0}-Y_{0,0}\right)+E\left(Y_{0,1}-Y_{0,0}\right)\right]=E\left(Y_{1,1}-Y_{0,1}-Y_{1,0}+Y_{0,0}\right) \tag{2}$$

We assume that both treatments are randomly assigned and independent of each other such that the different ATEs are identified as

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

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

2.2 Long and short regression models

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

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

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

The fully saturated "long" 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; see Appendix A.2 for detailed derivations.[7] The coefficients in the long regression model correspond to the main effects of T_1 and T_2 against a

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

business-as-usual counterfactual and the interaction effect:

$$\beta_1 = E\left(Y_{1,0} - Y_{0,0}\right), \tag{5}$$
$$\beta_2 = E\left(Y_{0,1} - Y_{0,0}\right), \tag{6}$$
$$\beta_{12} = E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right). \tag{7}$$

By contrast, the regression coefficients in the short model are

$$\beta_1^s = E\left(Y_{1,1} - Y_{0,1}\right) P\left(T_2 = 1\right) + E\left(Y_{1,0} - Y_{0,0}\right) P\left(T_2 = 0\right) \tag{8}$$
$$= E\left(Y_{1,0} - Y_{0,0}\right) + E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right) P(T_2 = 1) \tag{9}$$
$$= \beta_1 + \beta_{12} P(T_2 = 1)$$

and

$$\beta_2^s = E\left(Y_{1,1} - Y_{1,0}\right) P\left(T_1 = 1\right) + E\left(Y_{0,1} - Y_{0,0}\right) P\left(T_1 = 0\right) \tag{10}$$
$$= E\left(Y_{0,1} - Y_{0,0}\right) + E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right) P(T_1 = 1) \tag{11}$$
$$= \beta_2 + \beta_{12} P(T_1 = 1)$$

Equation (8) shows that β_1^s yields a weighted average of the ATE of T_1 relative to a counterfactual where $T_2 = 1$ and the ATE of T_1 relative to a business-as-usual counterfactual where $T_2 = 0$. The weights correspond to the fractions of individuals with $T_2 = 1$ and $T_2 = 0$, which are determined by the experimental design. Alternatively, β_1^s can be written as 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 $T_2 = 1$; see Equation (9). Equations (10) and (11) present the corresponding expressions for β_2^s.

These derivations show that, unless the interaction effect is zero (in which case $\beta_1 = \beta_1^s$ and $\beta_2 = \beta_2^s$), the population regression coefficients in the short regression model neither correspond to the main effects nor the interaction effect. Instead, the short model yields a composite treatment effect that is a weighted average of ATEs relative to different counterfactuals.[8]

[8] From a theoretical perspective, the choice between the long and the short model is related to the problem of making inference on a single treatment effect with covariates, where one has to decide whether to include the covariates linearly and to make inference on a weighted average of treatment effects or to run fully saturated regressions and to make inference on the average treatment effects (e.g., Angrist & Krueger, 1999; Angrist & Pischke, 2009). However, the practical implications are not the same because experimental treatments are fundamentally different in nature from standard covariates; see Section 2.3 for a discussion.

2.3 Long or short model: What do we care about?

Section 2.2 shows that the long model identifies the main effects relative to a business-as-usual counterfactual, whereas the short model yields a weighted average of treatment effects that depends on the nature and distribution of the other treatment arms in the experiment. However, this weighted average is typically neither of primary academic interest nor policy-relevant. This view is consistent with how papers we reanalyze motivate their object of interest, which is usually the main treatment effect against a business-as-usual counterfactual. Of the 19 papers in Table A.1 in Appendix A.1 that present results from the short model without all interactions, we did not find any study that mentioned (in the main text or in a footnote) that the presented treatment effects should be interpreted as either (a) a composite effect that includes a weighted average of the interaction with the other treatments or (b) as being against a counterfactual that was not business-as-usual but one that also had the other treatments in the same experiment.

One way to make the case for the short model is to recast the problem we identify as one of external rather than internal validity. Specifically, all experiments are carried out in a context with several unobserved covariates and thus any experimental treatment effect is a weighted average of the treatment interacted with a distribution of unobserved covariates. If the other experimental arms are considered as analogous to unobserved covariates, then inference on treatment effects based on the short model can be considered internally valid. In this view, the challenge is that the unobserved covariates (including other treatment arms) will vary across contexts.

However, experimental treatments are fundamentally different in nature from standard covariates. They are determined by the experimenter based on research interest, and rarely represent real-world counterfactuals. For example, in some cases, the interventions studied are new and the other treatments may not even exist in the study population. Even if they do exist, there is no reason to believe that the distributions in the population mirror those in the experiment. Thus, we view this issue as a challenge to internal validity because the other experimental arms are also *controlled by the researcher* and not just a set of "background unobservable factors". Further, papers with factorial designs often use the two-step procedure described in Section 4, 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.

There are settings where focusing on the short model may be fine. For example, in experiments that focus on testing concepts or establishing existence results (such as resume audit studies to study discrimination), treatment estimates are unlikely to directly affect discussions about policy or program implementation, and a weighted average effect may be a reasonable target parameter, subject to researchers indicating clearly how the resulting effect should be interpreted.

However, even in settings where the coefficients in the short model are of interest, they can always be constructed based on the coefficients in the long model, while the converse is not true. One can also use the long model to test hypotheses about the coefficients in the short regression model: $H_0 : \beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1) = 0$. Which test is more powerful depends on the relative magnitude of the four experimental cells.[9] Unlike the short model, the long model additionally allows for testing a rich variety of hypotheses about counterfactual effects such as $H_0 : \beta_1 + \beta_{12} p = 0$ for policy-relevant values of p, which generally differ from the experimental assignment probability $P(T_2 = 1)$.

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

2.4 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. In what follows, we focus on β_1. The analysis for β_2 is symmetric and omitted.

Under random assignment and standard regularity conditions, the OLS estimator of β_1 based on the long regression model, $\hat{\beta}_1$, is consistent:

$$\hat{\beta}_1 \overset{p}{\to} \beta_1 = E\left(Y_{1,0} - Y_{0,0}\right)$$

By contrast, the probability limit of the OLS estimators based on the short model is

$$\hat{\beta}_1^s \overset{p}{\to} \beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1).$$

Unless the true interaction effect is zero (i.e., $\beta_{12} = 0$), $\hat{\beta}_1^s$ is not consistent for the main effects relative to a business-as-usual counterfactual. Thus, if the goal is to achieve consistency for the main effects, one should always use the long model.

The choice between the long and the short regression model is less clear cut when it

[9] For example, when $N_1 = N_2 = N_3 = N_4 = N/4$, the tests based on the long model and the short model exhibit the same power. In practice, we recommend comparing both tests when doing power calculations.

comes to inference. To illustrate, suppose that the data generating process is given by

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

where $\varepsilon_i \sim N(0, \sigma^2)$ is independent of (T_{1i}, T_{2i}) and σ^2 is known. 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.

If the interaction effect is zero (i.e., $\beta_{12} = 0$), it follows from standard results that, conditional on $(T_{11}, \ldots, T_{1N}, T_{21}, \ldots, T_{2N})$,

$$\hat{\beta}_1 \sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right) \text{ and } \hat{\beta}_1^s \sim N\left(\beta_1, Var\left(\hat{\beta}_1^s\right)\right),$$

where

$$Var\left(\hat{\beta}_1^s\right) \leq Var\left(\hat{\beta}_1\right).$$

As a consequence, the t-test based on the short model exhibits higher finite sample power than the t-test based on the long model. Appendix A.3 gives explicit formulas of $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_1)$ in terms of (N_1, N_2, N_3, N_4), provides a formal comparison between the power of the long and the short model, and discusses the role of the "size" of the interaction cell, N_4.

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 3.1. Depending on the true value of the interaction effect, the finite sample power of the t-test based on the short model can be higher or lower than the power of the t-test 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.[10] Of the 3,505 articles published in this period 124 (3.5%) are field experiments (Table A.1 provides more details). Factorial designs are widely used: Among 124 field experiments 27 (22%) had a factorial design.[11] Only

[10]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".

[11]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

8 of these 27 articles with factorial designs (∼30%) used the long model including all interaction terms as their main specification (see Table 2).

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

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

3.1 Ignoring the interaction: Theory

The discussion above highlights that it is common for experimental papers with factorial designs to ignore the interaction and focus on the short regression model. This is theoretically justified if the researcher is certain that all the interactions are zero, in which case it leads to consistent estimates of the main effects and to power improvements relative to the long model (see Section 2.4). However, if the interactions are not zero, ignoring the interaction yields inconsistent estimates and size distortions.

To illustrate, we introduce a running example based on a prototypical setting to which we will return throughout the paper. We focus on the problem of testing the null hypothesis that the main effect of T_1 is equal to zero, $H_0 : \beta_1 = 0$. The analysis for β_2 is symmetric and omitted. We consider a 2×2 design with a total sample size of $N = 1,000$, where $N_1 = N_2 = N_3 = N_4 = 250$. The data are generated as

$$Y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} T_{2i} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1),$$

where T_{1i} and T_{2i} are randomly assigned treatments with $P(T_{1i} = 1) = P(T_{2i} = 1) = 0.5$. This experiment has power 90% to detect an effect size of 0.2σ at the 5% level using

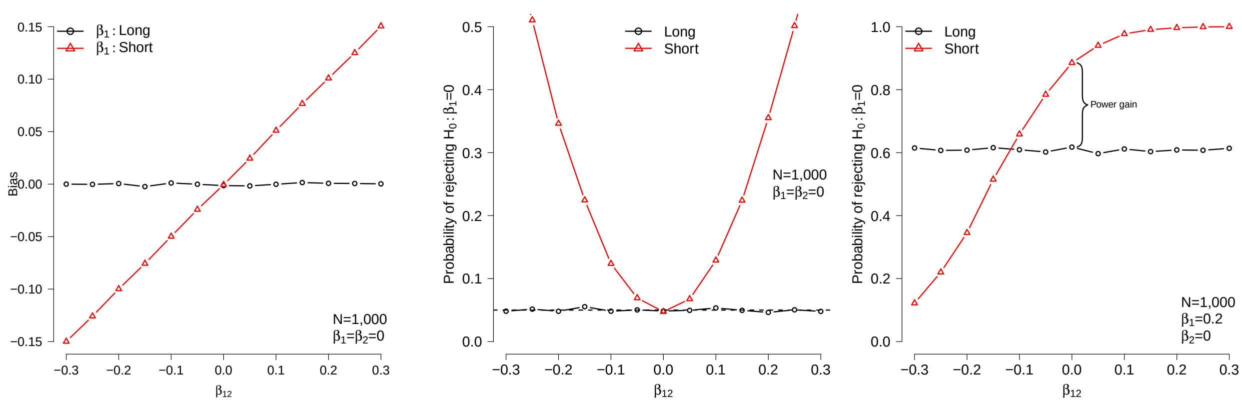
re-randomized to study behavioral changes conditional on a realization of the previous treatment. Examples of studies with two-stage randomization designs include Cohen & Dupas (2010), Karlan & Zinman (2009), and Ashraf et al. (2010). Finally, we do not include experiments where there is no "treatment", but rather conditions are randomized to elicit individuals preference parameters (e.g., Andersen et al., 2008; Gneezy et al., 2009; Fisman et al., 2008).

the short regression.[12] We use Monte Carlo simulations to assess the rejection rates of different inference procedures under the null (size) and the alternative hypothesis (power).

Figure 1 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 whenever $\beta_{12} \neq 0$. The trade-off is that for $\beta_{12} = 0$, the short model controls size and exhibits higher 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 1 is that researchers should avoid the short model, unless there is no uncertainty that $\beta_{12} = 0$.

Figure 1: 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 1b and 1c is $\alpha = 0.05$.

3.2 Ignoring the interaction: Practice

Here, we examine the practical implications of ignoring the interactions in the papers listed in Table A.1. 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, 19 papers do

[12] The minimum detectable effect for the long model with power 90% and size 5% is 0.29σ.

not include all interaction terms. Of these 19, 4 papers did not have publicly-available replication data.[13]

We downloaded the publicly-available data files and replicated the main results in each of the remaining 15 papers. We standardized the outcome variable in each paper to have mean zero and standard deviation of one. We then compared the original treatment effects (estimated without the interaction terms) with those estimated including the interaction terms. In other words, we compare estimates based on the short model (Equation (4)) to those based on the long model (Equation (3)).

3.2.1 Key facts about interactions

As the discussion above highlights, the extent to which the short model will not control size depends on the value of the interactions in practice. We therefore start by plotting the distribution of estimated interaction effects (Figure 2) and documenting facts regarding interactions from our re-analysis. We find that interactions are quantitatively important and typically not second-order. While the median (mean) interaction for these papers is 0.00σ (0.00σ), the median (mean) absolute value of the interaction is 0.07σ (0.13σ). The median (mean) absolute value of interactions relative to the main treatment effects is 0.37 (1.55). Thus, while it may be true that interactions are small on average across all studies, they are often sizeable in any given study. As we discuss below, this leads to a considerable extent of incorrect inference in any given study when interactions are not included.

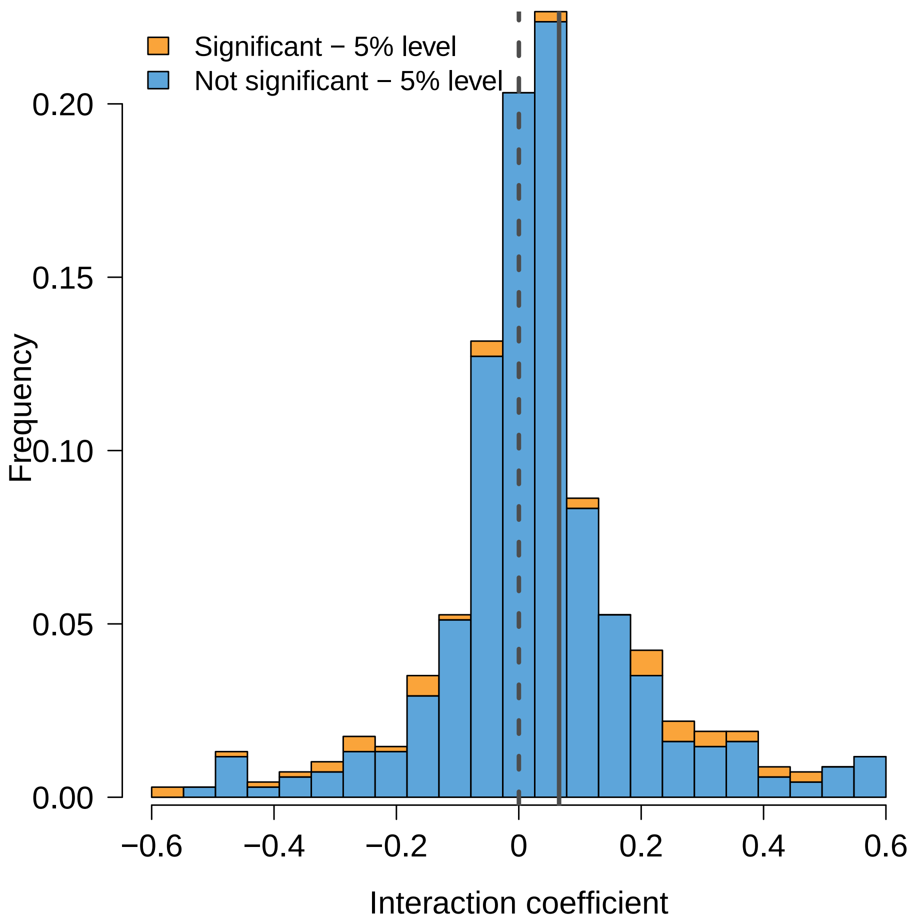
The second key finding is that despite the interactions being quantitatively important, most experiments will rarely reject the null hypothesis that they are zero (Figure 2 shades the fraction of the interactions that are significant in the studies that we re-analyze). Among the 15 papers that we re-analyzed, 6.2% of interactions are significant at the 10% level, 3.6% are significant at the 5% level, and 0.9% are significant at the 1% level.[14] These findings are not surprising because factorial designs are *rarely powered* to detect meaningful interactions. For example, Gelman (2018) shows that one 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 the main effects. Thus, the lack of inclusion of interactions may reflect authors' beliefs that the interactions are second order as inferred from their lack of significance in the long model.

[13] Online Appendix B.1 (in http://mauricio-romero.com/pdfs/papers/Appendix_crosscuts.pdf) describes the experimental design of the 27 papers.

[14] Among the papers that originally included all interactions, 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0.0% are significant at the 1% level.

The implication of these results is that it is rarely justified to implement a factorial design with the aim of *detecting* interactions since most experiments are not powered for this. Rather, the reason for factorial designs seems to be increasing power to detect the main effects. However, as we show below, this comes at the considerable cost of an increased rate of false positives (which is unsurprising based on the distribution of interactions shown in Figure 2).

Figure 2: Distribution of the interactions



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

3.2.2 Implications of ignoring interactions

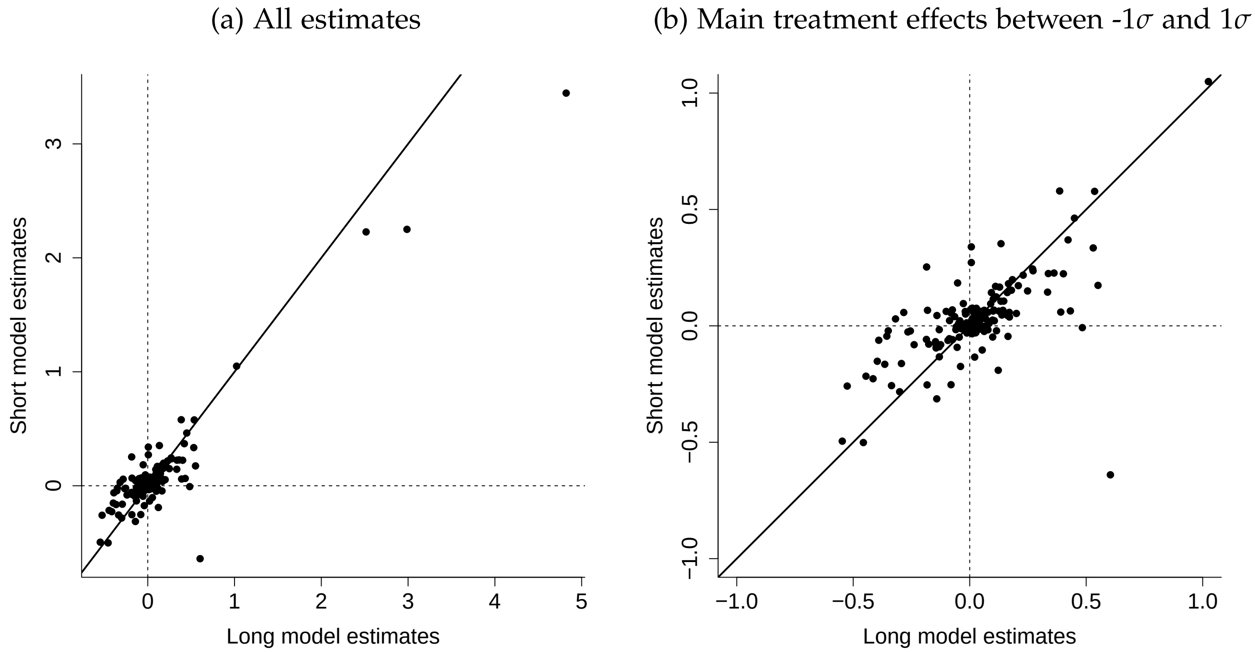
Figure 3a compares the original treatment effect estimates based on the short model to the estimates based on the long model which includes the interaction terms (Figure 3b zooms in to cases where the value of the main treatment effects in the short model is between -1 to 1 standard deviation). The median change in the absolute value of the point estimate of the main treatment effect is 96%. Roughly 26% of estimated treatment effects change sign when they are estimated using the long regression.

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

We find similar results 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.1.2 (Figure A.2 and Table A.2).

Finally, following the discussion in Section 2.3, we also distinguish between policy and conceptual experiments in Table A.1 (the latter typically have more treatments and interactions) and see that the problem of incorrect inference from ignoring interaction terms remains even when we restrict attention to the policy experiments. Of the 12 policy experiments, 9 do not include all interactions. When we re-estimate the treatment effects in these 9 papers after including all interactions, we find that out of 19 results that were significant at the 5% level in the paper, 6 (or 32%) are no longer so after including all interactions. Corresponding figures and tables are presented in Appendix A.1.3 (Figure A.4 and Table A.3).

Figure 3: Treatment estimates based on the long and the short model



Note: This figure shows how the main treatment estimates change between the short and the long model across all studies. Figure 3a has all the treatment effects, while Figure 3b zooms in to cases where the value of the main treatment effects in the short model is between -1 to 1 standard deviation. The median main treatment estimate from the short model is 0.01σ, the median main treatment estimate from the long model is 0.02σ, the average absolute difference between the treatment estimates of the short and the long model is 0.05σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 96%, and 26% of treatment estimates change sign when they are estimated using the long or the short model.

Table 3: Significance of treatment estimates based on the long and the short model

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

This table shows the number of 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 Model selection (or pre-testing) yields invalid inferences

As implied by the quote from Kremer (2003), researchers often recognize that using the short model is only correct for inference on the main treatment effect if the interaction is close to zero. However, the problem is that the value of the interaction is not known *ex ante* (also see the discussion in Section 5.3). 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 two-sided t-test.

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

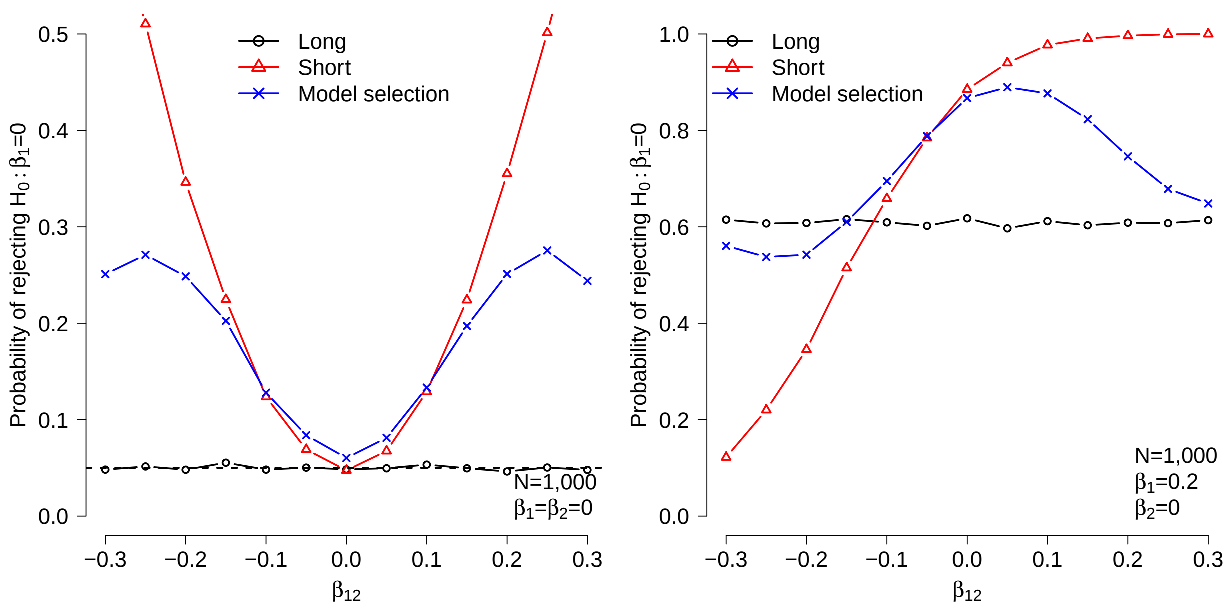
   (b) If $H_0 : \beta_{12} = 0$ is not rejected, test $H_0 : \beta_1 = 0$ using the two-sided t-test based on the short model.

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). To illustrate this issue, we return to our running example. The size and power properties of the two-step model selection approach are shown in Figure 4. For reference, we also include results for the t-tests based on the long and the short model. The main takeaway from Figure 4 is that model selection leads to incorrect inferences and false positives. Thus, researchers should always avoid it.

The performance of the model selection approach to determine whether one should run the short or the long model is particularly poor because field experiments are rarely powered to reject that the interactions are zero. In our running example, the power to detect interactions in the range of values of $-0.2\sigma < \beta_{12} < 0.2\sigma$ is 61% at most (at the 5% level). These simulation-based results are also borne out in practice. Figure 2 shows that only 3.6% of interactions were significant at the 5% level in our re-analysis. Thus, using a rejection threshold of 5%, the model-selection approach would lead to estimating the short model in over 96% of the cases we re-analyze. Thus, the rate of incorrect inference under model-selection will continue to be nearly as high as just running the short model.

Figure 4: Model selection does not control for size

(a) Size (b) Power



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.*

5 Can we improve power while achieving size control?

The motivation for factorial designs and estimating the short model is often the belief that interactions are "small". The problem in practice is that the actual value of the interaction is not known *ex ante* and both the common approaches of directly estimating the short model or doing a two-step model selection procedure do not control size. We now examine whether it is possible to improve power relative to t-tests based on the long model, while maintaining size control for relevant values of the interactions. We consider 2×2 factorial designs and refer to Section 7 for a discussion of factorial designs with more than two treatments.

To simplify the exposition, we focus on β_1 and partial out T_2, keeping the partialling-out implicit. The analysis for β_2 is symmetric and omitted. Defining $T_{12} \equiv T_1 T_2$, the regression model of interest is[15]

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

[15] We omit the intercept because all variables have mean zero after partialling-out T_2 and the constant.

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

5.1 Optimality properties of the *t*-test based on the long model

The two-sided *t*-test based on the long regression model is the uniformly most powerful test among tests that are unbiased for all values of the interaction effect (e.g., van der Vaart, 1998).[16] The practical implication of this classical result is that any procedure that is more powerful than the *t*-test for some values of the interaction must underperform somewhere else. As a consequence, to achieve higher power than the *t*-test based on the long model, one has to make a choice about which values of the interaction to direct power to. In practice, this choice needs to be made based on some form of prior knowledge.

Even if one is willing to direct power to particular values of the interaction and to sacrifice power somewhere else, the scope for power improvements relative the two-sided *t*-test based on the long regression model is limited if one insists on uniform size control. The reason is that for the corresponding one-sided testing problem, the usual one-sided *t*-test based on the long model is the uniformly most powerful test among all tests (e.g., Proposition 15.2 in van der Vaart, 1998). Thus, at any parameter value, the uniformly most powerful test is a one-sided *t*-test and the best one can hope for is to improve the power from the two-sided to a one-sided test (see, e.g., Armstrong et al. (2019) and Armstrong & Kolesar (2019) for a further discussion of this point). For 5% tests, this power improvement is never larger than 12.5 percentage points. It can also be shown that the scope for improving the average length of the usual confidence intervals based on the long regression model is limited (e.g., Armstrong & Kolesar, 2018, 2019; Armstrong et al., 2019).[17]

Section 5.2 proposes a nearly optimal test which comes close to achieving the maximal power gain at a priori likely values of the interaction, while controlling size for all values of the interaction. In Section 5.3, we explore an approach based on prior knowledge on the magnitude of the interaction. We show that when the prior knowledge is correct, this approach controls size and yields power gains relative to the *t*-test based on the long model. However, unlike the *t*-test based on the long model and the nearly optimal test, it suffers from size distortions if the prior knowledge is incorrect. Appendix A.5 explores two additional econometric approaches based on work by Imbens & Manski (2004), Stoye (2009) and McCloskey (2017).

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

[17] Moreover, the results in Joshi (1969) imply the usual two-sided confidence interval based on the long regression model achieves minimax expected length (Armstrong & Kolesar, 2019).

5.2 Nearly optimal tests targeting power towards a likely value $\bar{\beta}_{12}$

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

$$H_0: \ \beta_1 = 0, \ \ \beta_{12} \in \mathbb{R} \quad \text{against} \quad H_1: \ \beta_1 \neq 0, \ \ \beta_{12} = 0. \tag{13}$$

We use the numerical algorithm developed by Elliott et al. (2015) to construct a nearly optimal test for the testing problem (13). To describe their procedure, note that under standard conditions, the t-statistics are approximately normally distributed in large samples

$$\begin{pmatrix} \hat{t}_1 \\ \hat{t}_{12} \end{pmatrix} \sim N\left(\begin{pmatrix} t_1 \\ t_{12} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{14}$$

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).

Consider the problem of maximizing power in the following hypothesis testing problem:

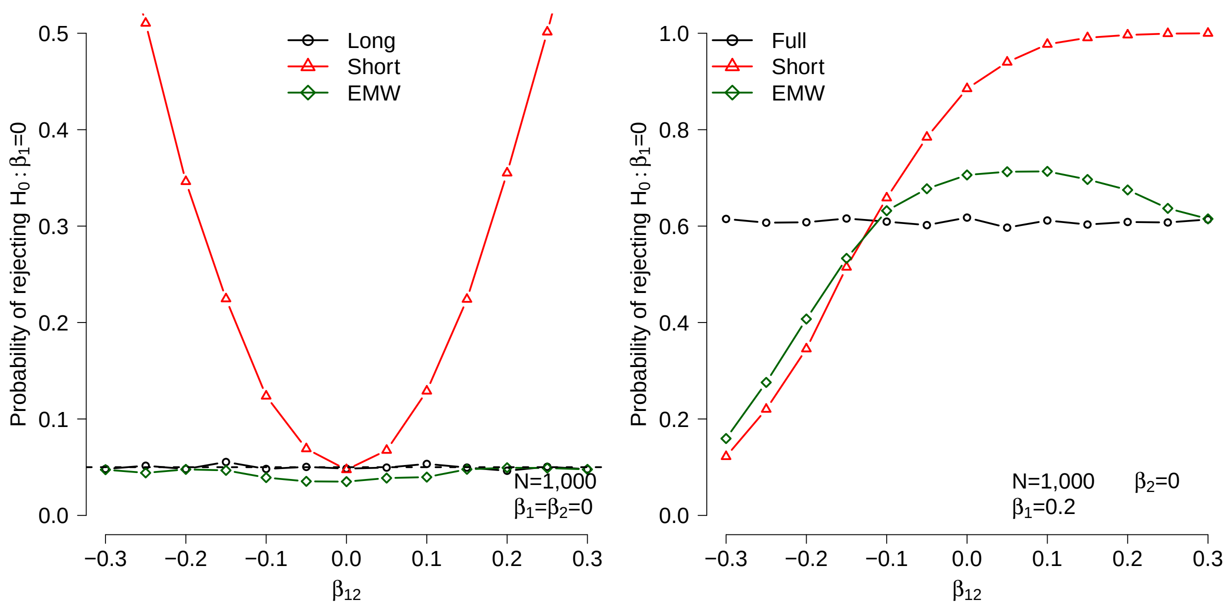
$$H_0: \ t_1 = 0, \ \ t_{12} \in \mathbb{R} \quad \text{against} \quad H_1: \ t_1 \neq 0, \ \ t_{12} = 0. \tag{15}$$

A common approach to construct powerful tests for problems with composite hypotheses is to choose tests based on their weighted average power. In particular, we seek 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 $\int f_t dF(t)$", where the weighting function F is chosen by the researcher. Now suppose that the null is replaced by "$H_{0,\Lambda}$: the density of $\hat{t}$ is $\int f_t d\Lambda(t)$". To obtain the best test, one needs 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).

Since it is generally difficult to analytically determine and computationally approximate Λ^{LF}, Elliott et al. (2015) 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 yields power gains over running the full model for "intermediate" values of β_{12}

(a) Size (b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 5a and 5b is $\alpha = 0.05$*. EMW refers to Elliott et al. (2015)'s nearly optimal test.*

Figure 5 displays the results of applying the nearly optimal test in the context of our running example.[18] The test controls size for all values of β_{12} and, by construction, is nearly optimal when $\beta_{12} = 0$. A comparison with the t-test based on the long model shows that the nearly optimal test is more powerful when β_{12} is close to zero. The nearly optimal test comes close to achieving the maximal possible power.[19] However, as expected given the discussion in Section 5.1, these power gains come at a cost. For certain values of β_{12}, the power can be much lower than that of the t-test based on the long model. Appendix A.6.3 provides a more comprehensive assessment of the performance of the nearly optimal tests by plotting power curves for different values of β_1.

[18]To improve the performance of their procedure, Elliott et al. (2015) 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.

[19]For example, when $\beta_1 = 0.1$ ($\beta_1 = 0.2$) the power of the nearly optimal is 78.3% (97.7%) of the power of the one-sided t-test.

5.3 Inference under a priori restrictions on the magnitude of β_{12}

Suppose that the researcher is certain that $\beta_{12} = \bar{\beta}_{12}$, in which case she can obtain powerful tests based on a regression of $Y - \bar{\beta}_{12}T_{12}$ on T_1. If $\bar{\beta}_{12} = 0$, this corresponds to estimating the short model. As shown in Section 2.4, the t-test based on the short model is more powerful than t-test based on the long model when the prior knowledge that $\beta_{12} = 0$ is correct, but does not control size when it is not.

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

Assumption 1. $|\beta_{12}| \leq C$ *for some finite constant* C.

Assumption 1 restricts the parameter space for β_{12} and implies that

$$\beta_{12} \in \{b_{12} : |b_{12}| \leq C\} \equiv \mathcal{B}_{12}.$$

Here we use the approach developed in Armstrong & Kolesar (2018) and Armstrong et al. (2019) to construct optimal confidence intervals under Assumption 1.[20] To describe their procedure, we write model (12) in matrix form as

$$\mathbf{Y} = \beta_1 \mathbf{T}_1 + \beta_{12}\mathbf{T}_{12} + \varepsilon \tag{16}$$

and assume that $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_N)$ and that σ^2 is known. The implementation with heteroskedastic and non-Gaussian errors is discussed in Appendix A.4. An affine estimator of β_1 can be written as $\hat{\beta}_1 = a + b'\mathbf{Y}$, for some a and b that can depend on $\mathbf{X} \equiv (\mathbf{T}_1, \mathbf{T}_{12})$. For example, for the long OLS regression model, $a = 0$ and b is the first row of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Define the "worst case" biases as

$$\begin{aligned}
\overline{\text{Bias}}\left(\hat{\beta}_1\right) &= \sup_{\beta_1 \in \mathbb{R}, \beta_{12} \in \mathcal{B}_{12}} E_{(\beta_1, \beta_{12})}\left(\hat{\beta}_1 - \beta_1\right), \\
\underline{\text{Bias}}\left(\hat{\beta}_1\right) &= \inf_{\beta_1 \in \mathbb{R}, \beta_{12} \in \mathcal{B}_{12}} E_{(\beta_1, \beta_{12})}\left(\hat{\beta}_1 - \beta_1\right),
\end{aligned}$$

where $E_{(\beta_1, \beta_{12})}$ denotes the expectation under the distribution generated by model (16) with (β_1, β_{12}). Assuming that $(\mathbf{T}_1, \mathbf{T}_{12})$ are fixed, $\hat{\beta}_1$ is normally distributed with mean $a + b'(\beta_1 \mathbf{T}_1 + \beta_{12}\mathbf{T}_{12})$ and variance $SE(\hat{\beta}_1)^2 = \|b\|_2^2 \sigma^2$. Thus, as (β_1, β_{12}) varies over

[20] Optimality here refers to minimizing the width of the confidence intervals. We focus on the width of the confidence intervals because of the intuitive appeal and practical relevance of this criterion. If one were to optimize the power of the test that the confidence interval inverts, the resulting procedure would be different in general.

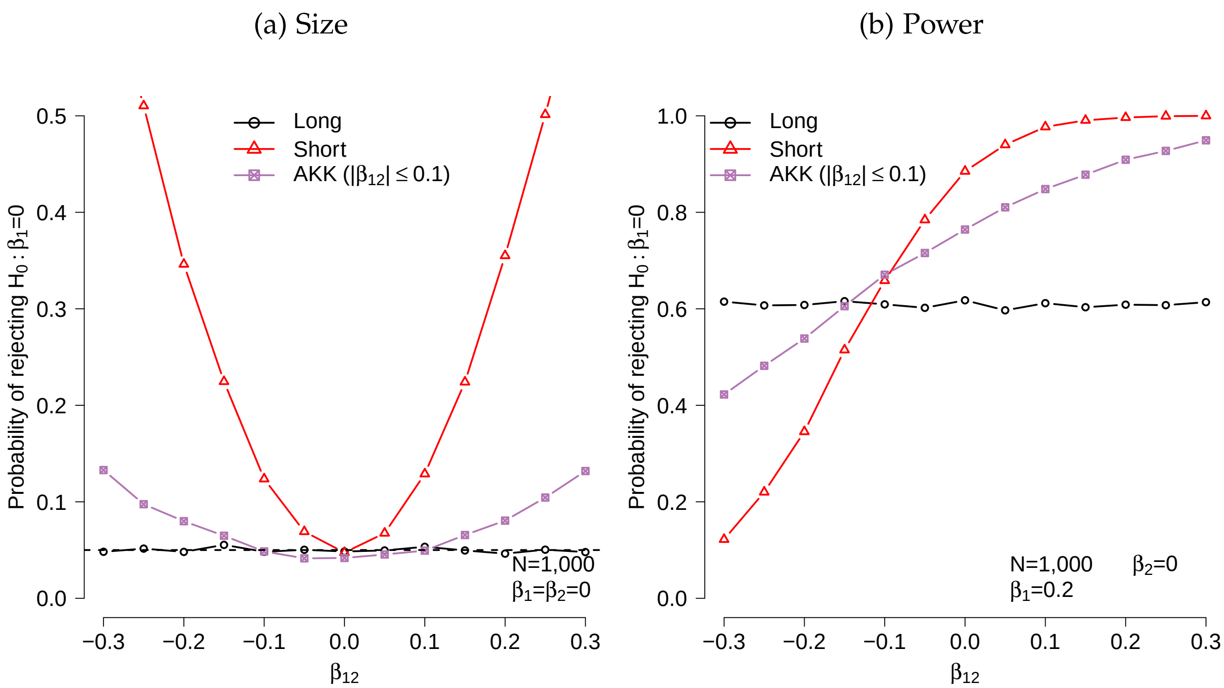
$\mathbb{R} \times \mathcal{B}_{12}$, the t-ratio, $\frac{(\hat{\beta}_1 - \beta_1)}{SE(\hat{\beta}_1)}$, is normally distributed with variance one and mean varying from $\frac{\underline{\text{Bias}}(\hat{\beta}_1)}{SE(\hat{\beta}_1)}$ to $\frac{\overline{\text{Bias}}(\hat{\beta}_1)}{SE(\hat{\beta}_1)}$. To construct a two-sided confidence interval, note that testing $H_0 : \beta_1 = \beta_1^0$ based on a t-statistic with critical value $cv_\alpha\left(\frac{\max\{|\underline{\text{Bias}}(\hat{\beta}_1)|, |\overline{\text{Bias}}(\hat{\beta}_1)|\}}{SE(\hat{\beta}_1)}\right)$ yields a level α test, where $cv_\alpha(t)$ denotes the $1 - \alpha$ quantile of a folded normal distribution with location parameter t and scale parameter 1. Inverting this test yields the following confidence interval:

$$\hat{\beta}_1 \pm cv_\alpha\left(\frac{\max\left\{|\underline{\text{Bias}}\left(\hat{\beta}_1\right)|, |\overline{\text{Bias}}\left(\hat{\beta}_1\right)|\right\}}{SE(\hat{\beta}_1)}\right) SE(\hat{\beta}_1) \tag{17}$$

The length of the confidence interval (17) is determined by the bias and the variance of the estimator $\hat{\beta}_1$, and to obtain optimal confidence intervals one has to solve a bias-variance trade-off. This problem is amenable to convex optimization and we describe how to solve it in Appendix A.4.

Figure 6 reports the rejection probabilities of a test that rejects if zero is not in the confidence interval. For the purpose of illustration, we consider $C = 0.1$ such that $\mathcal{B}_{12} = [-0.1, 0.1]$. Our results suggest that imposing prior knowledge in the form of an upper bound on the magnitude of the interaction effect can yield substantial power improvements relative to the t-tests based on the long regression model, while controlling size when this prior knowledge is in fact correct. However, this method exhibits size distortions when the prior knowledge is incorrect, i.e., when $|\beta_{12}| > 0.1$. Appendix A.6.4 presents the corresponding power curves for different values of β_1.

Figure 6: Restrictions on the magnitude of β_{12} yield power gains if they are correct but lead to incorrect inferences if they are not



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 6a and 6b is $\alpha = 0.05$*. AKK refers to Armstrong et al. (2019) approach for constructing optimal confidence intervals under prior knowledge about the magnitude of* β_{12}*.*

6 Should we run experiments with factorial designs?

The discussion above focused on improving inference in existing experiments with factorial designs. However, for the design of new experiments, a natural question is: Should we run experiments with factorial designs in the first place? A design-based 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 such that

		T_1	
		No	*Yes*
T_2	*No*	N_1^*	N_2^*
	Yes	N_3^*	0

Consider the following regression model

$$Y = \beta_0^* + \beta_1^* T_1 + \beta_2^* T_2 + \varepsilon^*. \tag{18}$$

Let $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ denote the OLS estimators of β_1^* and β_2^*. We show in Appendix A.2.3 that if T_1 and T_2 are randomly assigned, $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are consistent for the respective main effects.

To illustrate the power implications of leaving the interaction cell empty, consider an experiment where the researcher cares equally about power to detect an effect of T_1 and T_2, and thus assigns the same sample size to both treatments: $N_2^* = N_3^* = N_T^*$. In what follows, we focus on β_1^*. The analysis for β_2^* is symmetric and omitted. Under the assumptions of Section A.3.1, the (conditional) variance of $\hat{\beta}_1^*$ is given by

$$Var\left(\hat{\beta}_1^*\right) = \sigma^2 \frac{N - N_T^*}{(N - 2N_T^*)N_T^*}.$$

$Var\left(\hat{\beta}_1^*\right)$ is minimized when $N_T^* = \frac{N}{2}\left(2 - \sqrt{2}\right)$ and we assume that the experiment is designed in this manner.[21] A comparison to the variance of the estimator based on the long model, $\hat{\beta}_1$, shows that $Var\left(\hat{\beta}_1^*\right) \leq Var\left(\hat{\beta}_1\right)$.[22] Thus, by the same reasoning as in Section 2.4, leaving the interaction cell empty leads to power improvements for testing hypotheses about the main effects 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 the interaction. Moreover, among the approaches that achieve size control for all values of β_{12} (the long model and the nearly optimal test), leaving the interaction cell empty yields the highest power. This design (with the interaction cells empty) yields power gains relative to running two separate experiments, because the control group is used twice. But it avoids the problem of interactions discussed above.

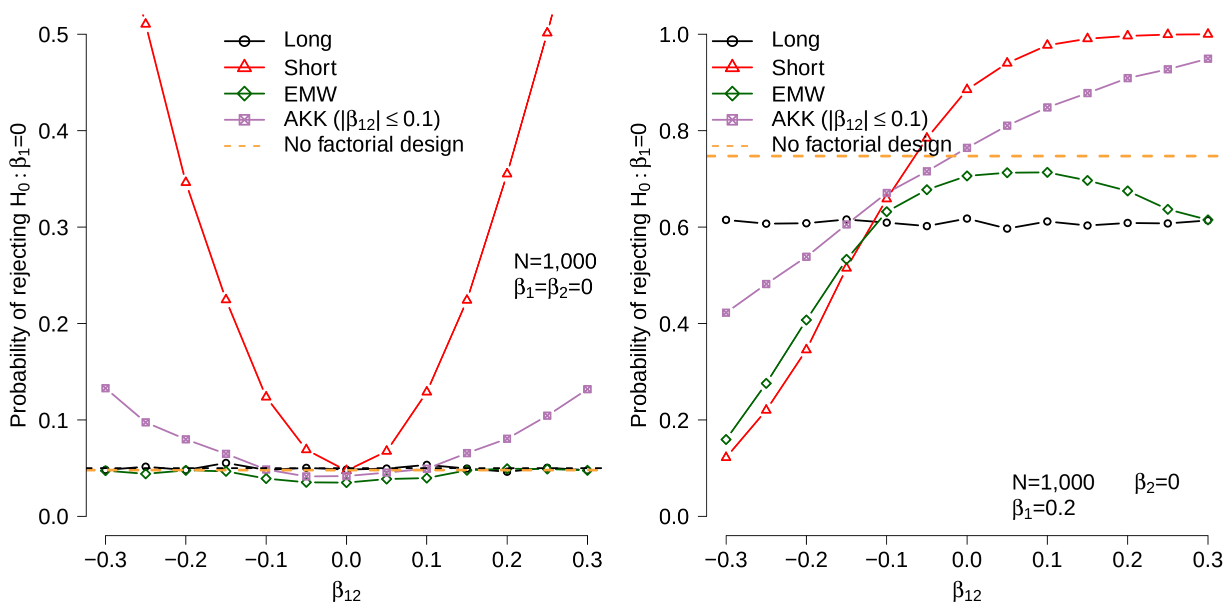
Thus, if one is not interested in interaction effects, we recommend avoiding factorial designs and leaving the interaction cell empty. An example of such a design is provided by Muralidharan & Sundararaman (2011) who study the impact of four different interventions in one experiment with one common control group, but no cross-cutting treatment arms.

[21] This exact sample split is impossible in any application since $\frac{N}{2}\left(2 - \sqrt{2}\right)$ is not an integer. In our simulations we therefore use $N_T^* = 0.29N$ and $N_1^* = 0.42N$.

[22] For this comparison, we assume that both experiments are designed such that they exhibit equal power to detect an effect of T_1 and T_2.

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

(a) Size (b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 7a and 7b is $\alpha = 0.05$*. EMW refers to Elliott et al. (2015)'s nearly optimal test. AKK refers to Armstrong et al. (2019) approach for constructing optimal confidence intervals under prior knowledge about the magnitude of* β_{12}*.*

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.

The theoretical analysis of Section 2 straightforwardly extends to more complicated factorial designs. In particular, estimators based on the long regression model are consistent for the main and interaction effects, whereas the estimators based on the short regression model are consistent for weighted averages of treatment effects with respect to the counterfactuals defined by the other arms of the experiment. The more treatments there are, the more complicated the interpretation of these composite effects will be. It is thus not surprising that *none* of the experimental papers with high-dimensional designs in our reanalysis that ignore the interactions motivate their experiment as being about estimating these composite effects.

Conceptually, both 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. Incorporating prior knowledge in the form of restrictions on the magnitude of interactions can be problematic in practice because this approach requires reliable prior knowledge on the magnitude of potentially very many interactions to yield notable power improvements.[23]

Therefore, our recommendation for inference in more complicated factorial designs is to use two-sided *t*-tests based on the long model. These tests are easy to compute irrespective of the dimensionality of the problem and have desirable optimality properties. When the interaction effects are not of primary interest, we recommend leaving the interaction cells empty at the design stage, which yields power improvements over the *t*-test based on the long model.

8 Discussion and conclusion

In this paper, we study the theory and practice of inference in factorial designs. We document that the popular approaches of directly estimating the short model or doing a two-step model selection procedure yield invalid inferences about the main effects. In contrast, the long model yields consistent estimates and always controls size. In practice, factorial designs are often motivated by the belief that the interactions are "small" relative to the main effects. We therefore explore whether it may be possible to increase power relative to the long model when the interactions are likely to be small. We show that local power gains near a priori likely small values of the interactions are possible, but that the scope for power improvements is small if one insists on size control for all values of the interactions. Thus, our recommendation for the analysis of completed experiments is to use the long regression model.

For the design of new experiments, an alternative is to leave the interaction cells empty and to increase the number of units assigned exclusively to one of the treatments or the control group. This simple design-based approach naturally controls size and yields notable global power improvements relative to the long model. We recommend this approach for policy experiments where a business-as-usual counterfactual is important.

Reviewing classic texts on experimental design, and reflecting on the historical use of factorial designs in field experiments, we identify three cases where the short model may be fine. The first is the case of iterative high-frequency experiments, where the goal of initial experiments is to explore several treatment dimensions in an efficient way

[23] Both approaches discussed in Appendix A.5 are computationally feasible in more complicated cross-cut designs.

to generate promising interventions for further testing. For example, Cochran & Cox (1957, p.152) recommend factorial designs for "exploratory work where the object is to determine quickly the effects of a number of factors over a specified range".

The second case is to improve an experiment's external validity. Cochran & Cox (1957, p.152) recommend factorial designs for "experiments designed to lead to recommendations that must apply over a wide range of conditions. Subsidiary factors may be brought into an experiment so as to test the principal factors under a variety of conditions similar to those that will be encountered in the population to which recommendations are to apply"; see also the discussion in Fisher (1992). This point may also be relevant to examples such as resume audit studies, where the characteristics being experimentally manipulated (such as age, education, gender, race, and work experience) also exist and vary in the population. Our discussion suggests that it may make sense for these studies to assign probabilities to specific characteristics that are similar to their incidence in the population. Doing so will make the estimated short-model coefficient more likely to approximate the population treatment effect of interest.

The third case is when the goal is not hypothesis testing but to minimize mean squared error (MSE) criteria (or other loss functions) which involve a bias-variance trade-off. For example, Blair et al. (2019) document that for small values of the interaction effects, estimators based on the short model can yield a lower root MSE than the estimators based on the design which leaves the interaction cell empty. Such alternative criteria are particularly relevant for problems where the goal is to make better decisions in the specific setting of the experiment. This is often the case for agricultural experiments that need to vary soil, moisture, temperature, fertilizer, and several other inputs to determine the ideal combination of input use. In these settings, the goal of the experiment is less about testing whether any of these factors have a "significant" effect, and more to make better-informed decisions regarding the optimal combination in which to use various inputs. Thus, it is not surprising that factorial designs were popular for agricultural experiments on test plots (e.g., Cochran & Cox, 1957).

However, policy experiments are expensive and difficult to run iteratively, and are typically used based on whether the intervention had a "significant" effect. Publication-bias towards significant findings is well documented, and evidence is often aggregated for policy by counting studies where an intervention has been found to be effective at conventional significance levels. Thus, the sensitivity of the significance of treatment effects to the inclusion/exclusion of interaction terms (as shown in this paper) is likely to have non-trivial implications for how evidence is summarized and translated into policy. Our recommendations for practice are motivated primarily by this concern.

In this paper, we focus on frequentist inference which is the most prevalent inference paradigm in experimental economics. However, especially in settings with many treatments, Bayesian hierarchical methods may constitute a useful framework for efficient learning in experiments with cross-cutting designs by adding additional parametric structure and prior knowledge (e.g., Kassler et al., 2019).

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.

Andrews, I., & Kasy, M. (2018). Identification of and correction for publication bias. *forthcoming American Economic Review*.

Angrist, J. D., & Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), (Vol. 3, p. 1277 - 1366). Elsevier. doi: https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics an empiricist's companion*. Princeton University Press.

Armstrong, T. B., & Kolesar, M. (2018). Optimal inference in a class of regression models. *Econometrica*, *86*(2), 655-683. doi: 10.3982/ECTA14434

Armstrong, T. B., & Kolesar, M. (2019). *Sensitivity analysis using approximate moment condition models.* arXiv:1808.07387.

Armstrong, T. B., Kolesar, M., & Kwon, S. (2019). *Optimal inference in regularized regression models.* Unpublished Manuscript.

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

Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, *113*(3), 838?859. doi: 10.1017/S0003055419000194

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.

Christensen, G., & Miguel, E. (2018, September). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920-80. doi: 10.1257/jel.20171350

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.

Fisher, R. A. (1992). The arrangement of field experiments. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 82–91). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4612-4380-9_8 doi: 10.1007/978-1-4612-4380-9_8

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.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. doi: 10.1126/science.1255484

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.

Joshi, V. M. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, *40*(3), 1042–1067.

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.

Kassler, D., Nichols-Barrer, I., & Finucane, M. (2019). Beyond treatment versus control: How bayesian analysis makes factorial experiments feasible in education research. *Evaluation Review*.

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. (2018, 11). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics*, *134*(2), 557-598. doi: 10.1093/qje/qjy029

A Appendix

A.1 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 July 4, 2019)

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Olken (2007)	Monitoring Corruption: Evidence from a Field Experiment in Indonesia	JPE	2007	1529	3	2	0	Yes	Yes
Banerjee et al. (2007)	Remedying Education: Evidence from Two Randomized Experiments in India	QJE	2007	1213	2	1	0	Yes	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	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	Yes
Karlan et al. (2014)	Agricultural Decisions after Relaxing Credit and Risk Constraints	QJE	2014	612	2	1	1	No	Yes
Bertrand et al. (2010)	What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment	QJE	2010	522	14	85	0	Yes	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	Policy Evaluation
Karlan & List (2007)	Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment	AER	2007	506	7	28	0	Yes	No
Thornton (2008)	The Demand for, and Impact of, Learning HIV Status	AER	2008	453	2	1	0	Yes	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	Yes
Alatas et al. (2012)	Targeting the Poor: Evidence from a Field Experiment in Indonesia	AER	2012	330	4	16	0	Yes	Yes
Karlan & Zinman (2008)	Credit Elasticities in Less-Developed Economies: Implications for Microfinance	AER	2008	311	3	2	0	Yes	No
Duflo et al. (2015a)	Education, HIV, and Early Fertility: Experimental Evidence from Kenya	AER	2015	282	3	3	1	Yes	Yes
Andreoni et al. (2017)	Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving	JPE	2017	270	2	1	1	Yes	No
Jakiela & Ozier (2015)	Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies	ReStud	2016	245	3	6	6	Yes	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	Policy Evaluation
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	No
Allcott & Taubinsky (2015)	Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market	AER	2015	237	2	1	0	No	No
Flory et al. (2014)	Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions	ReStud	2015	204	10	24	12	Yes	No
Brown et al. (2010)	Shrouded Attributes and Information Suppression: Evidence from the Field	QJE	2010	189	3	6	6	No	No
DellaVigna et al. (2016)	Voting to Tell Others	ReStud	2017	169	4	15	0	Yes	No
Fischer (2013)	Contract Structure, Risk-Sharing, and Investment Choice	ECMA	2013	162	7	9	9	Yes	No
Kaur et al. (2015)	Self-Control at Work	JPE	2015	154	8	16	0	Yes	No
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	Yes	Yes

Continued on next page

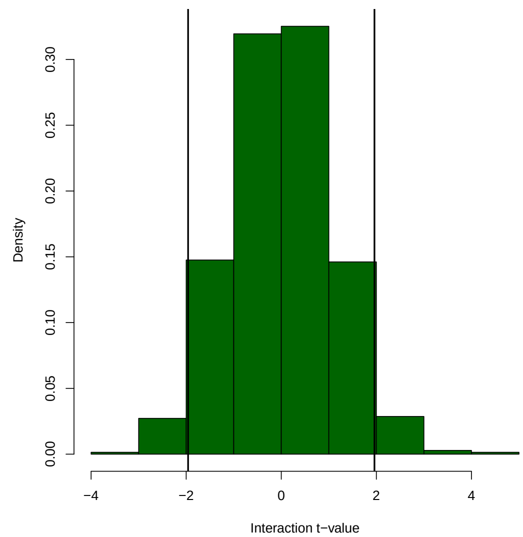
Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Blattman et al. (2017)	Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia	AER	2017	135	2	1	1	Yes	Yes
Khan et al. (2015)	Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors	QJE	2016	133	6	8	0	Yes	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	No
Kendall et al. (2015)	How Do Voters Respond to Information? Evidence from a Randomized Campaign	AER	2015	116	5	5	5	Yes	No
Pallais & Sands (2016)	Why the Referential Treatment? Evidence from Field Experiments on Referrals	JPE	2016	85	3	12	0	No	No

Note: This table provides relevant information from all articles with factorial designs published in top-5 journals. 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. Allcott & Taubinsky (2015) has two field experiments. The table refers to the second one. Section B.1.16 provides for more details. One of the three dimensions of randomization in Flory et al. (2014) does not appear in the publicly available data. Online Appendix B.1 (in http://mauricio-romero.com/pdfs/papers/Appendix_crosscuts.pdf) describes the experimental design of each of the 27 papers.

A.1.1 All papers

Figure A.1: Distribution of the *t*-value of interaction terms across studies



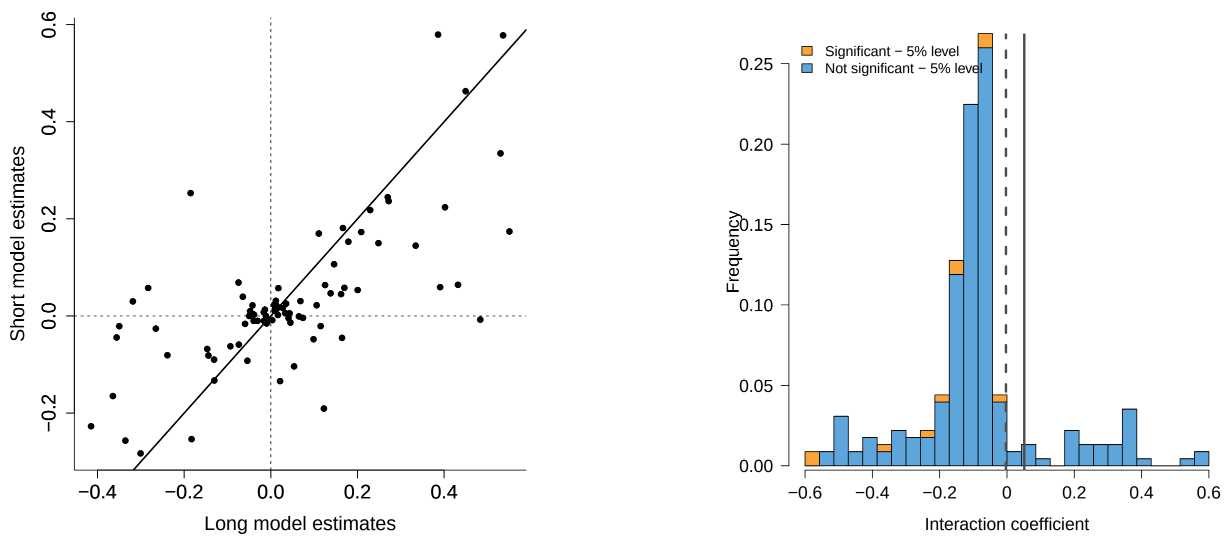
Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

A.1.2 Ten most cited papers

Figure A.2: Treatment estimates based on the long and the short model

(a) Main treatment estimates

(b) Interaction



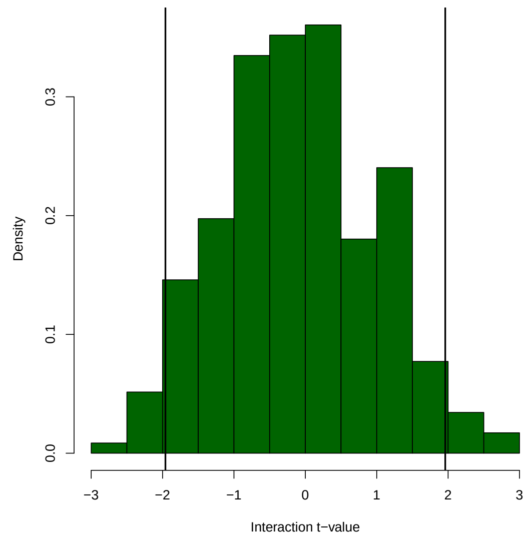
Note: Both figures show treatment estimates from the ten most cited papers with factorial designs and publicly available data that do not include the interaction in the original study. Figure A.2a shows how the main treatment estimates change across the short and the long model across studies. The median main treatment estimate from the short model is 0.01σ, the median main treatment estimate from the long model is 0.01σ, 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 131%, and 28% of treatment estimates change sign when they are estimated using the long instead of the short model. Figure A.2b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is -0.00σ (dashed vertical line), the median absolute value of the interactions is 0.05σ (dashed vertical line), 5.6% of interactions are significant at the 10% level, 2.6% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.37.

Table A.2: Significance of treatment estimates based on the long and the short model

Panel A: Significance at the 10% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	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.3: Distribution of the t-value of interaction terms across studies

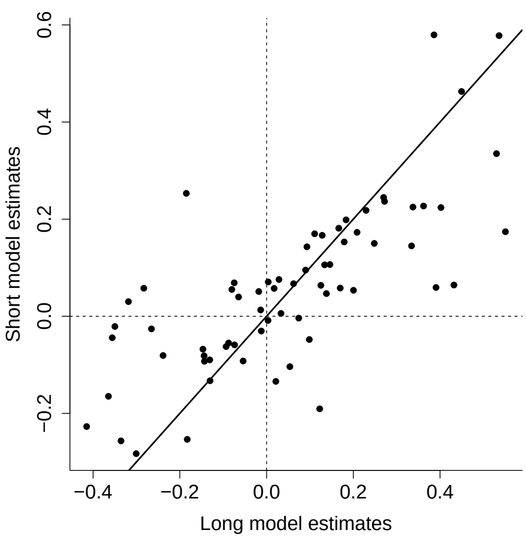


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

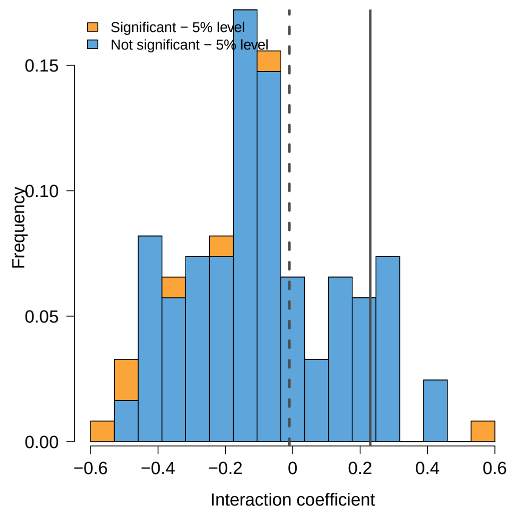
A.1.3 Policy experiments

Figure A.4: Treatment estimates from the long and the short regression

(a) Main treatment estimates



(b) Interaction



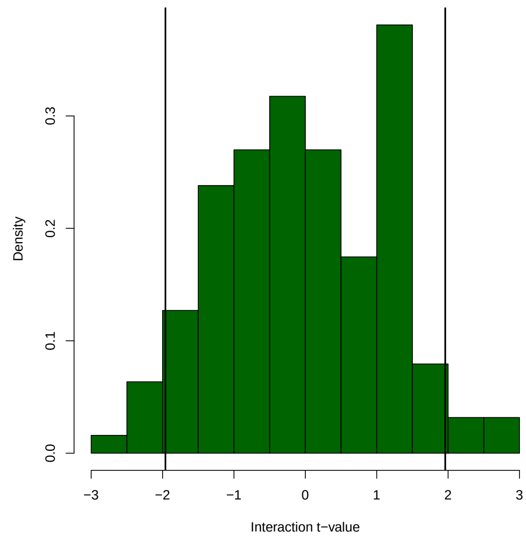
Note: Both figures show treatment estimates from the papers with factorial designs and publicly available data that do not include the interaction in the original study and do policy evaluation. Figure A.4a shows how the main treatment estimates change across the short and the long model across studies. The median main treatment estimate from the short model is 0.06σ, the median main treatment estimate from the long model is 0.05σ, the average absolute difference between the treatment estimates of the short and the long model is 0.07σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 69%, and 21% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.4b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is -0.01σ (dashed vertical line), the median absolute value of interactions is 0.23σ (solid vertical line), 6.3% of interactions are significant at the 10% level, 3.2% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 1.01.

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	31	10	41
Significant	5	21	26
Total	36	31	67
Panel B: Significance at the 5% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	43	6	49
Significant	5	13	18
Total	48	19	67
Panel C: Significance at the 1% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	56	3	59
Significant	1	7	8
Total	57	10	67

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 do policy evaluation. 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.5: Distribution of the t-value of interaction terms across studies

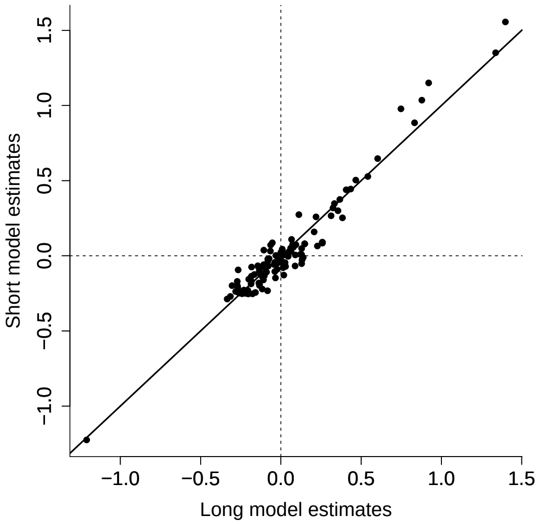


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

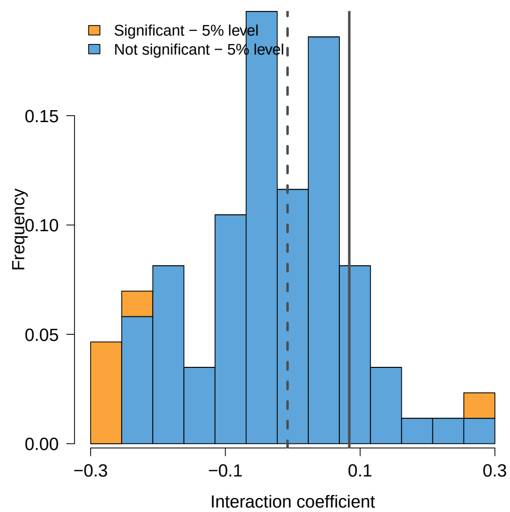
A.1.4 Studies with all interactions included

Figure A.6: Treatment estimates based on the long and the short model

(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 do policy evaluation. Figure A.6a 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 -0.03σ, the median main treatment estimate from the long model is -0.02σ, the average absolute difference between the treatment estimates of the short and the long model is 0.05σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 37%, and 15% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.6b shows the distribution of the interactions between the main treatments. We trim the top and bottom 1% of the distribution. The median interaction is -0.01σ (dashed vertical line), the median absolute value of interactions is 0.08σ (solid vertical line), 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.52.

Table A.4: Significance of treatment estimates based on the long and the short model

Panel A: Significance at the 10% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	61	13	74
Significant	4	39	43
Total	65	52	117

Panel B: Significance at the 5% level

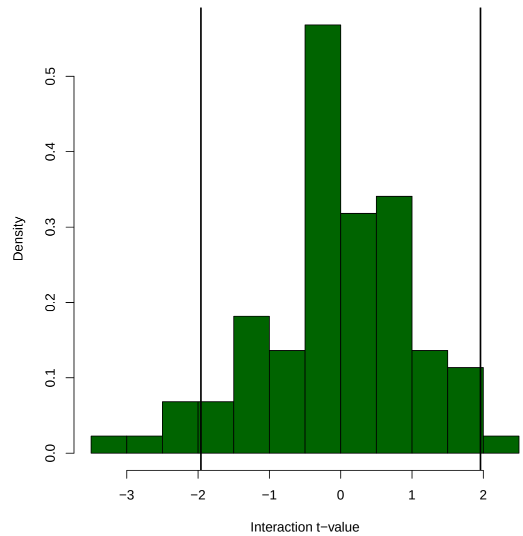
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	68	10	78
Significant	6	33	39
Total	74	43	117

Panel C: Significance at the 1% level

	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	77	12	89
Significant	2	26	28
Total	79	38	117

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 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.7: Distribution of the t-value of interaction terms across studies



Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96*.*

A.2 Derivation of expressions for the regression coefficients

A.2.1 Derivation of the expressions for β_1, β_2, and β_{12}

Because the long regression model (3) is fully saturated, we have

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

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

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

Thus, it follows that

$$\begin{aligned}\beta_1 &= E\left(Y_{1,0} - Y_{0,0}\right), \\ \beta_2 &= E\left(Y_{0,1} - Y_{0,0}\right), \\ \beta_{12} &= E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right).\end{aligned}$$

A.2.2 Derivation of the expressions for β_1^s and β_2^s

Here we derive (8). Equation (9) then follows from rearranging terms. The derivations of Equations (10) and (11) are similar and thus omitted.

For the short regression model (4), independence of T_1 and T_2 implies that

$$\beta_1^s = E\left(Y \mid T_1 = 1\right) - E\left(Y \mid T_1 = 0\right).$$

Consider

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

where the first equality follows from the law of iterated expectations and the second equality follows by the definition of potential outcomes and random assignment. Similarly, obtain

$$E\left(Y \mid T_1 = 0\right) = E\left(Y_{0,1}\right) P(T_2 = 1) + E\left(Y_{0,0}\right) P(T_2 = 0).$$

Thus, we have

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

A.2.3 Consistency of the OLS estimators based on model (18)

Here we show that when the interaction cell is empty and T_1 and T_2 are randomly assigned, the OLS estimators based on the regression model (18) are consistent for the main effects.

Define $\hat{\beta}^* \equiv (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)'$ and $\beta^* \equiv (\beta_0^*, \beta_1^*, \beta_2^*)' = E(XX')^{-1} E(XY)$, where $X = (1, T_1, T_2)'$. Under standard conditions, $\hat{\beta}^* \xrightarrow{p} \beta^*$. Hence, it remains to show that β_1^* and β_2^* are equal to the main effects. In what follows, we focus on β_1^*; the derivation for β_2^* is similar. To simplify the exposition, we define $p_1 \equiv P(T_1 = 1)$, $p_2 \equiv P(T_2 = 1)$ and $p_{12} \equiv P(T_1 = 1, T_2 = 1)$.

Multiplying out yields the following expressions for β_1^*:

$$\beta_1^* = \frac{(p_2 p_{12} - p_1 p_2)E(Y) + p_1(p_2 - p_2^2)E(Y \mid T_1 = 1) + p_2(p_1 p_2 - p_{12})E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2 + 2p_1 p_2 p_{12} - p_{12}^2}.$$

Using the fact that the interaction cell is empty, which implies that $p_{12} = 0$, obtain

$$\beta_1^* = \frac{-p_1 p_2 E(Y) + p_1 p_2(1 - p_2)E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \tag{19}$$

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

$$E(Y) = E(Y \mid T_1 = 1, T_2 = 0)p_1 + E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2) + E(Y \mid T_1 = 0, T_2 = 1)p_2. \tag{20}$$

Combining (19) and (20) and simplifying yields:

$$\beta_1^* = E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0)$$

The result now follows by random assignment of T_1 and T_2 and the definition of potential outcomes.

A.3 Variance reductions and power gains based on the short model

A.3.1 Formal power comparison between the short and the long model

Suppose that the researcher has access to a random sample $\{Y_i, T_{1i}, T_{2i}\}_{i=1}^N$ and that the data are generated according to the following linear model

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

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

Define $\boldsymbol{T}_1 \equiv (T_{11}, \ldots, T_{1N})'$ and $\boldsymbol{T}_2 \equiv (T_{21}, \ldots, T_{2N})'$. If the interaction effect is zero (i.e., $\beta_{12} = 0$), it follows from standard results that, conditional on $(\boldsymbol{T}_1, \boldsymbol{T}_2)$, $\hat{\beta}_1 \sim N\left(\beta_1, Var\left(\hat{\beta}_1\right)\right)$ and $\hat{\beta}_1^s \sim N\left(\beta_1, Var\left(\hat{\beta}_1^s\right)\right)$, where

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

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. We show that 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.[24]

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 $c_{1-\alpha/2} \equiv \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the quantile function of the standard normal distribution and $\alpha \in (0, 0.5)$ is the nominal significance level.

Lemma 1. *Suppose that the assumptions stated in the text hold and that $\beta_{12} = 0$. Then:*

(i) The finite sample power of the t-tests based on the short and the long model is 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:

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

[24] To see this, note that

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

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

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

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): To establish the result, we show that the power is decreasing in the standard error. 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

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

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[25]

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

which follows from the shape of the normal distribution. □

A.3.2 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 2.4 that, in a 2×2 factorial design, the variance of the estimate of β_1 is given by

$$
Var\left(\hat{\beta}_1\right) = \sigma^2\frac{N_1 + N_2}{N_1N_2} \quad \text{and} \quad Var\left(\hat{\beta}_1^s\right) = \sigma^2\frac{N_1N_3 + N_1N_4 + N_2N_3 + N_2N_4}{N_1N_2N_3 + N_1N_2N_4 + N_1N_3N_4 + N_2N_3N_4}.
$$

Moreover, as shown in Lemma 1, the power of the t-test is decreasing in the variance of the estimator.

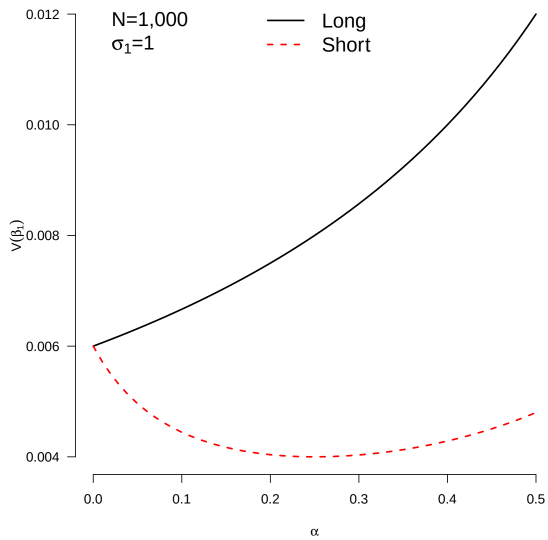
To illustrate, we simplify the problem by assuming that $N_1 = N_2 = N_3$, and hence that the researcher simply has to determine the relative size of N_4. Let α be such that $N_4 = \alpha N$. Thus, $N_1 = N_2 = N_3 = \frac{1}{3}(1-\alpha)N$. Then:

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

Figure A.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 lowest variance (highest power) 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.

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

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



A.4 Implementation details for Section 5.3

Recall that under Assumption 1, $\beta_{12} \in \{b_{12} : |b_{12}| \leq C\} \equiv \mathcal{B}_{12}$. Hence, our problem falls into the regularized regression setting of Armstrong et al. (2019). We therefore adopt the algorithm outlined in their Section 5 to our problem. The algorithm has three steps:[26]

1. Obtain an estimator $\hat{\sigma}^2$ of σ^2 by taking the square root of the average of the squared residuals from estimating the long model by OLS.

2. Minimize $cv_\alpha\left(\frac{|\text{Bias}(\hat{\beta}_\lambda)|}{\text{SE}(\hat{\beta}_\lambda)}\right)\text{SE}\left(\hat{\beta}_\lambda\right)$ with respect to λ over $[0, \infty)$, where

$$\text{SE}\left(\hat{\beta}_\lambda\right) \equiv \sqrt{\hat{\sigma}^2 \frac{\|\mathbf{T}_1 - \mathbf{T}_{12}\pi_\lambda\|_2^2}{\left((\mathbf{T}_1 - \mathbf{T}_{12}\pi_\lambda)'\,\mathbf{T}_1\right)^2}}$$

$$\text{Bias}\left(\hat{\beta}_\lambda\right) \equiv \frac{C}{|\pi_\lambda|}\frac{(\mathbf{T}_1 - \mathbf{T}_{12}\pi_\lambda)'\mathbf{T}_{12}\pi_\lambda}{(\mathbf{T}_1 - \mathbf{T}_{12}\pi_\lambda)'\mathbf{T}_1}$$

and π_λ solves $\min_\pi \|\mathbf{T}_1 - \pi\mathbf{T}_{12}\|_2^2 + \lambda|\pi|$. Denote the solution by λ^*.

[26]The implementation of the optimal confidence intervals with potentially heteroscedastic and non-Gaussian errors mimics the common practice of applying OLS (the validity of which requires homoscedasticity) in conjunction with heteroscedasticity robust standard errors, rather than weighted least squares.

3. Construct an optimal confidence interval as

$$\hat{\beta}_{\lambda^*} \pm cv_\alpha \left(\frac{|\text{Bias}\left(\hat{\beta}_{\lambda^*}\right)|}{\text{SE}\left(\hat{\beta}_{\lambda^*}\right)} \right) \text{SE}\left(\hat{\beta}_{\lambda^*}\right),$$

   where

$$\hat{\beta}_{\lambda^*} = \frac{\left(\mathbf{T}_1 - \mathbf{T}_2 \pi_{\lambda^*}\right)' \mathbf{Y}}{\left(\mathbf{T}_1 - \mathbf{T}_2 \pi_{\lambda^*}\right)' \mathbf{T}_1}.$$

   In this last step, we use the residuals from the initial estimate to construct a heteroscedasticity robust version of $\text{SE}\left(\hat{\beta}_{\lambda^*}\right)$.

A.5 Additional econometric approaches

In this section, we discuss two additional econometric approaches.

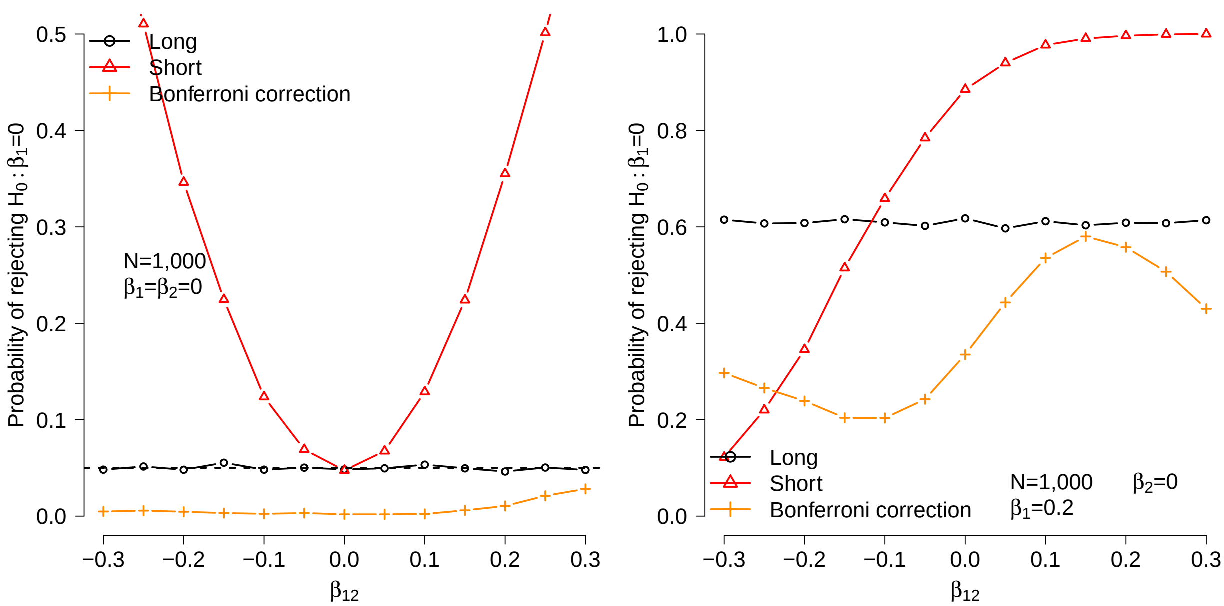
A.5.1 Model selection with a Bonferroni-style correction

A natural approach to control size in the presence of model selection is to take a least favorable (LF) approach and to use the largest critical value across all values for the nuisance parameter (e.g., D. W. K. 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.[27]

[27] 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 A.2: McCloskey (2017)'s Bonferroni-style correction controls size but does not exhibit power gains relative to the long model

(a) Size (b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures A.2a and A.2b 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 A.2 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 does not yield power gains relative to the t-test based on the long model, at least not over the regions of the parameter space considered here.[28]

A.5.2 An alternative inference approach based on Assumption 1

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

$$\beta \equiv (\beta_0, \beta_1, \beta_2)' = E\left(XX'\right)^{-1} E\left(X(Y - \beta_{12} T_{12})\right),$$

[28] 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. However, as we discuss in Section 5.1, the scope for improving power relative to the t-tests based on the long regression model is limited.

Assumption 1 implies that β_{12} lies in a compact interval,

$$\beta_{12} \in [-C, C] \equiv \left[\beta_{12}^{l}, \beta_{12}^{u}\right].$$

The population regression coefficient from a regression of $Y - \beta_{12}T_{12}$ on X is

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

Note that $E\left(XX'\right)^{-1} E\left(XT_{12}\right) \equiv (\gamma_0, \gamma_1, \gamma_2)'$ is the population regression coefficient from a regression of T_{12} on X. Independence of T_1 and T_2 implies that $\gamma_1 = E(T_{12} \mid T_1 = 1) - E(T_{12} \mid T_1 = 0)$ 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

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

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 can therefore apply the approach of Imbens & Manski (2004) and Stoye (2009) to construct confidence intervals for β_t:[29]

$$CI_{1-\alpha} = \left[\hat{\beta}_t^{l} - c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^{l}\right)}{N}},\ \hat{\beta}_t^{u} + c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^{u}\right)}{N}}\right], \tag{21}$$

where the critical value c_{IM} solves

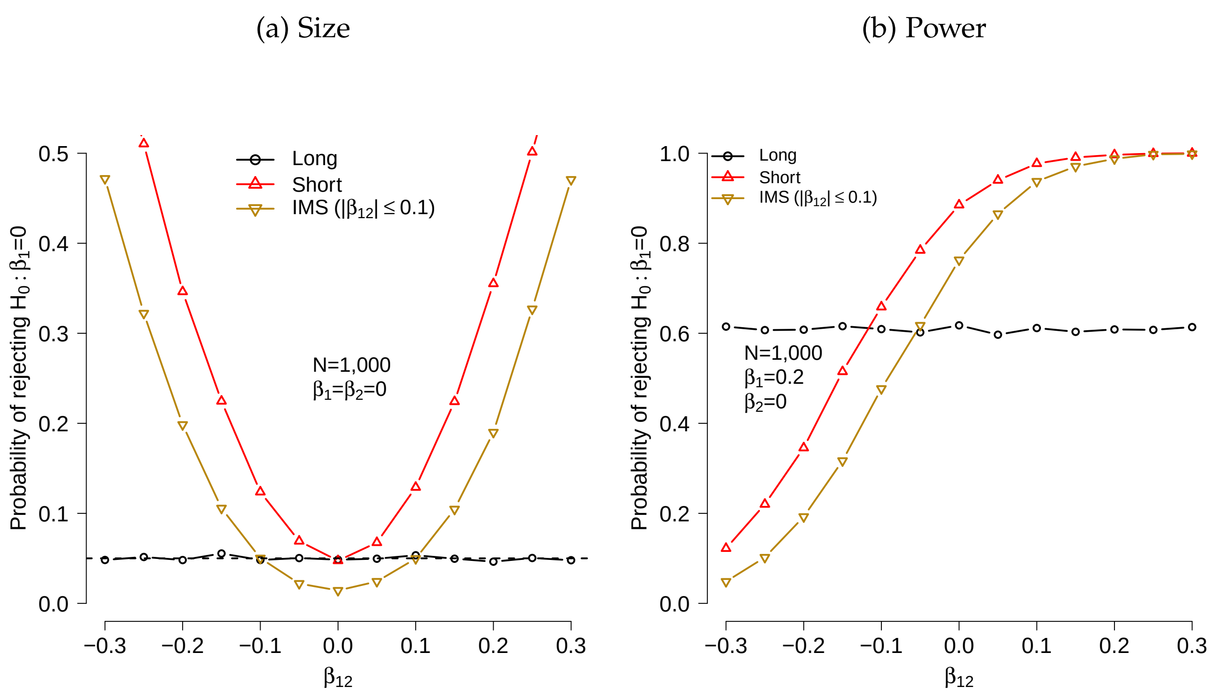
$$\Phi\left(c_{IM} + \sqrt{N} \cdot \frac{\hat{\beta}_t^{u} - \hat{\beta}_t^{l}}{\sqrt{\max\left(\widehat{Avar}\left(\hat{\beta}_t^{l}\right), \widehat{Avar}\left(\hat{\beta}_t^{l}\right)\right)}}\right) - \Phi\left(-c_{IM}\right) = 1 - \alpha.$$

Imbens & Manski (2004) and Stoye (2009) show that (21) is a valid confidence interval for β_t.

[29] 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).

In Figure A.3, we report the rejection probabilities of a test that rejects if zero is not in the confidence interval (21). For the purpose of illustration, we assume that $C = 0.1$ which implies that $\beta_{12} \in [-0.1, 0.1]$. 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 exhibits substantial size distortions when the prior knowledge is incorrect.

Figure A.3: Restrictions on the magnitude of β_{12} yield power gains if they are correct but lead to incorrect inferences if they are not



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures A.3a and A.3b is $\alpha = 0.05$. *IMS refers to Imbens & Manski (2004) and Stoye (2009) approach for constructing valid confidence intervals under prior knowledge about the magnitude of* β_{12}.

A.6 Additional figures and tables

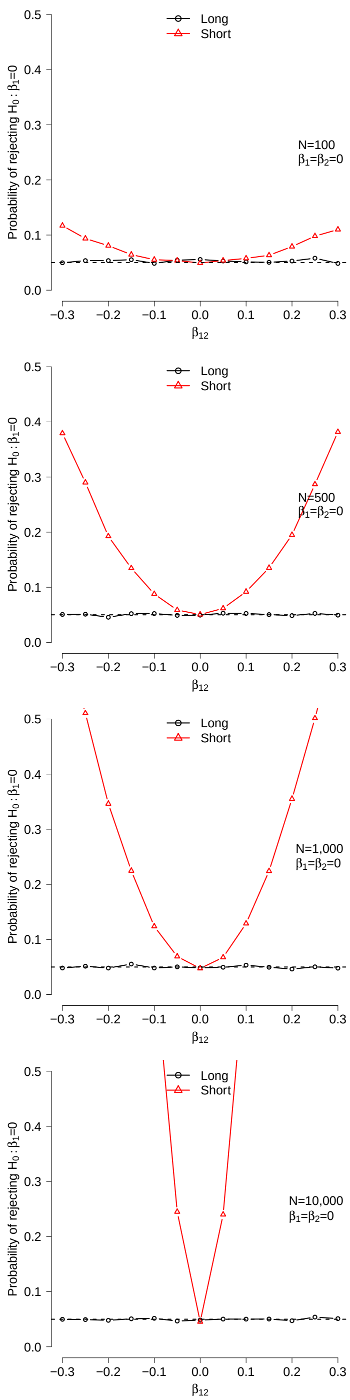
Table A.1: Articles publised in top-5 journals between 2006 and 2017

	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

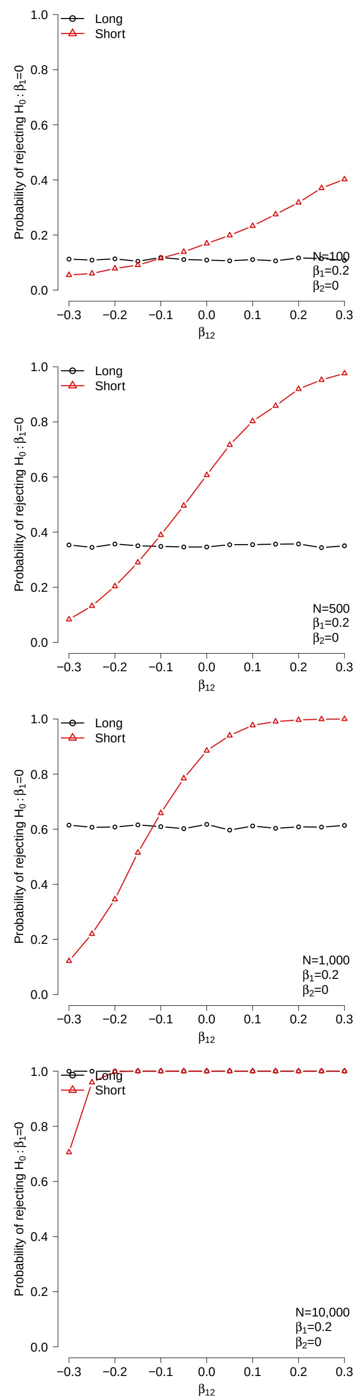
A.6.1 Ignoring the interaction

Figure A.1: Long and short model: Size and power

(a) Size

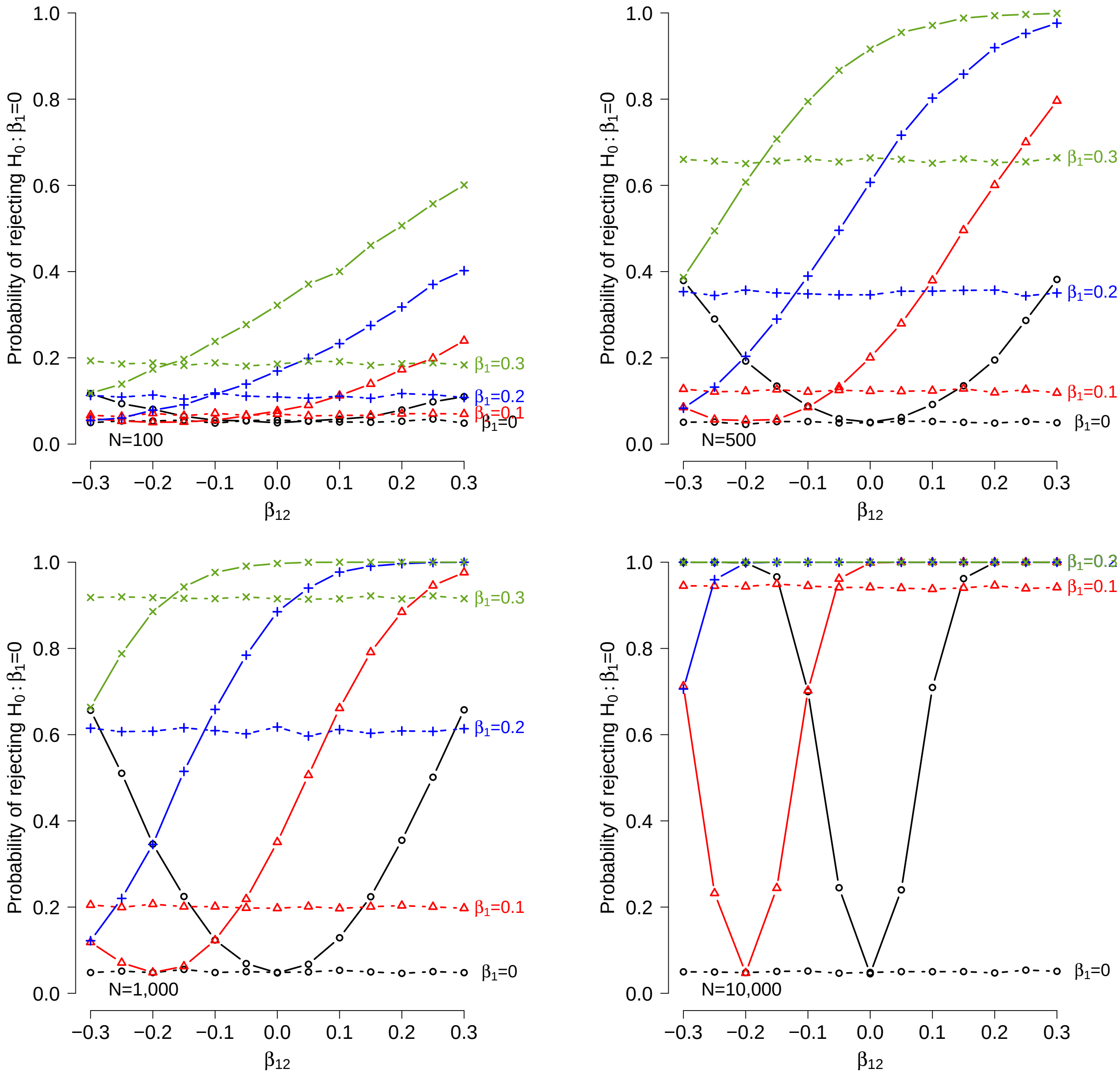


(b) Power



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

Figure A.2: Long and short model: Power curves

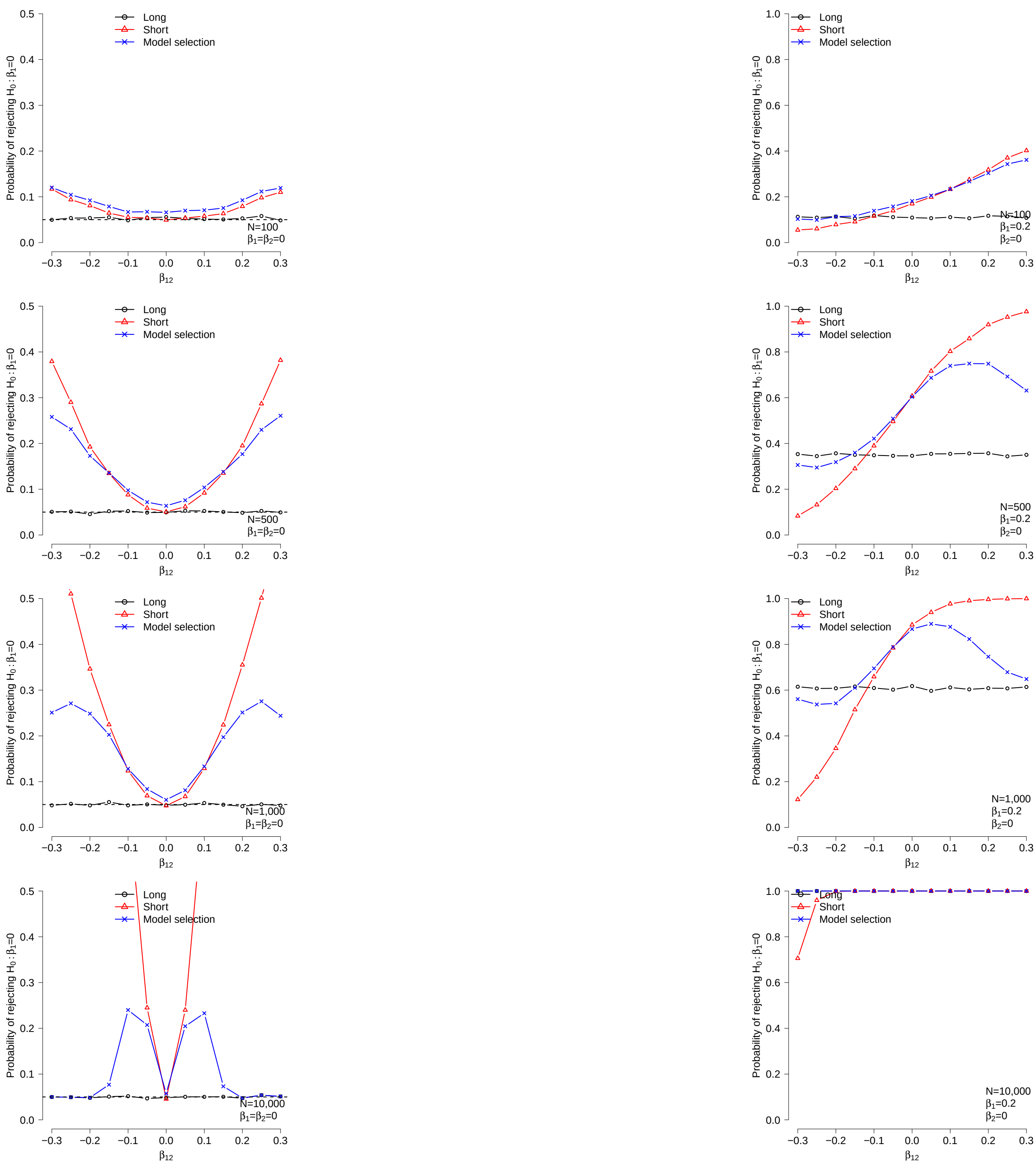


Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. In each figure, dashed lines show the power for the long model, while solid lines show power for the short model.

A.6.2 Pre-testing

Figure A.3: Model selection: Size and power

(a) Size (b) Power

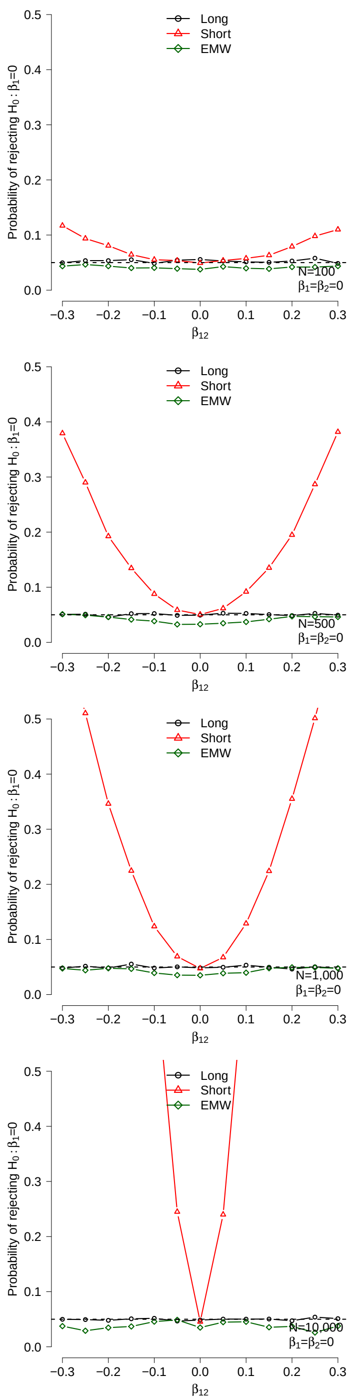


Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. *For the model selection, the short model is estimated if one fails to reject* $\beta_{12} = 0$ *at the 5% level.*

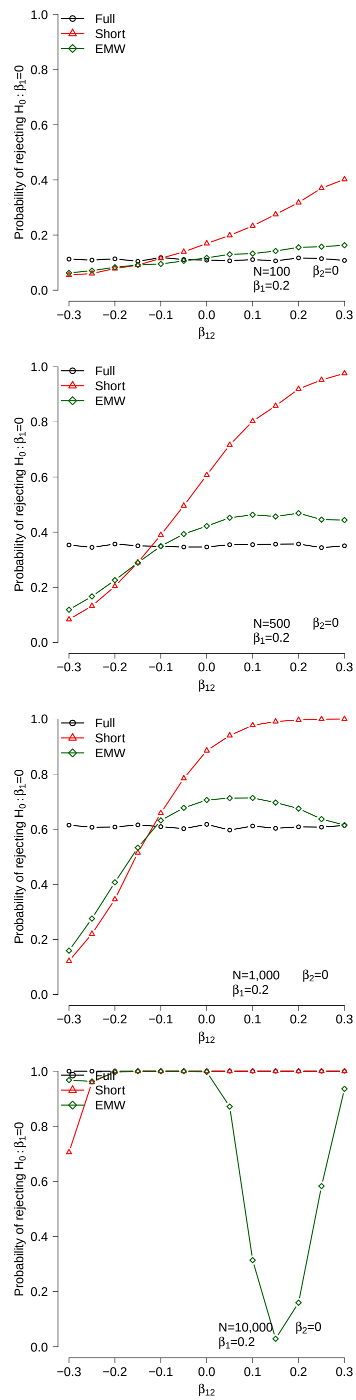
A.6.3 Elliott et al. (2015)'s nearly optimal test

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

(a) Size

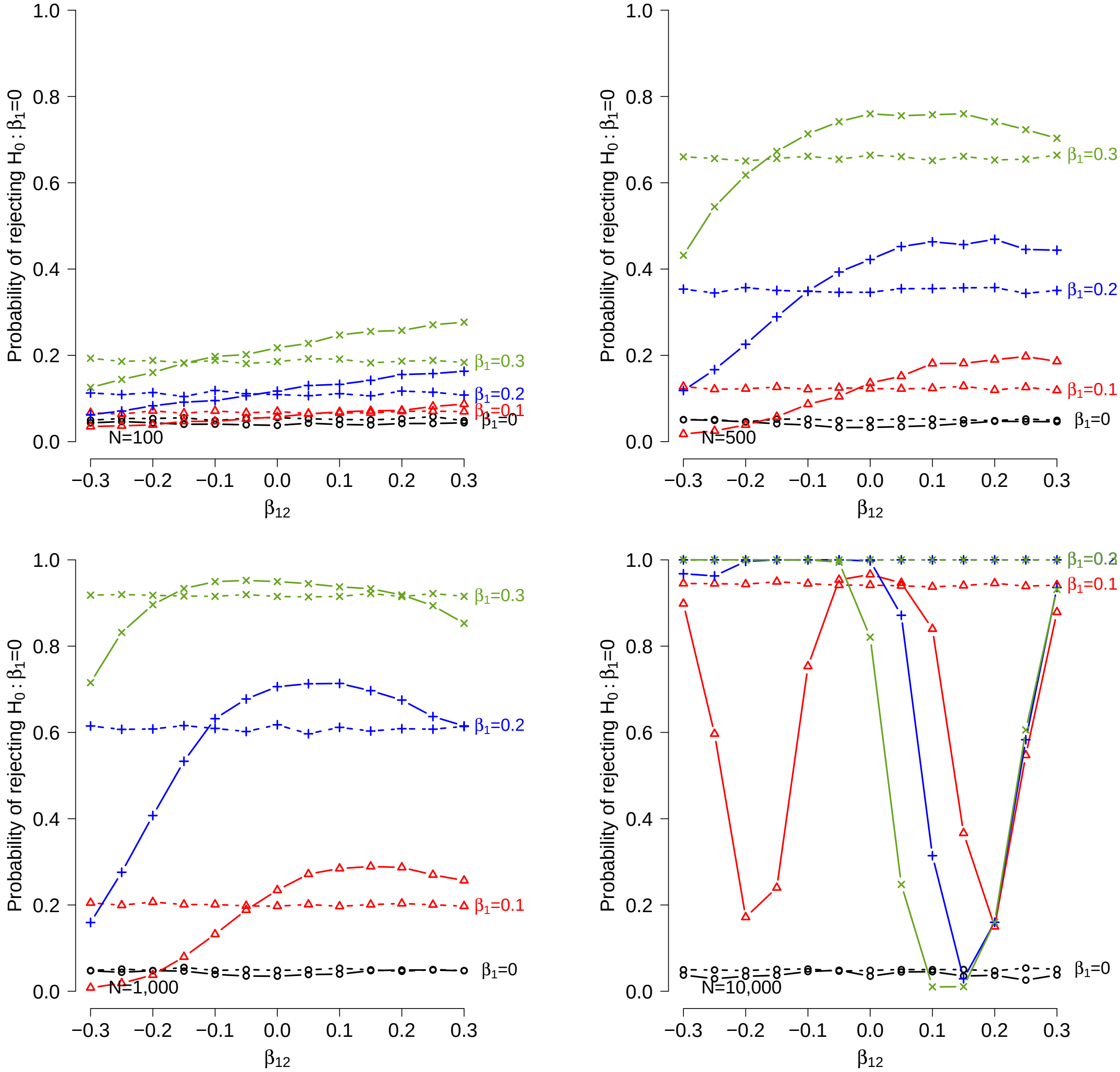


(b) Power



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

Figure A.5: Long model and Elliott et al. (2015)'s nearly optimal test: Power curves

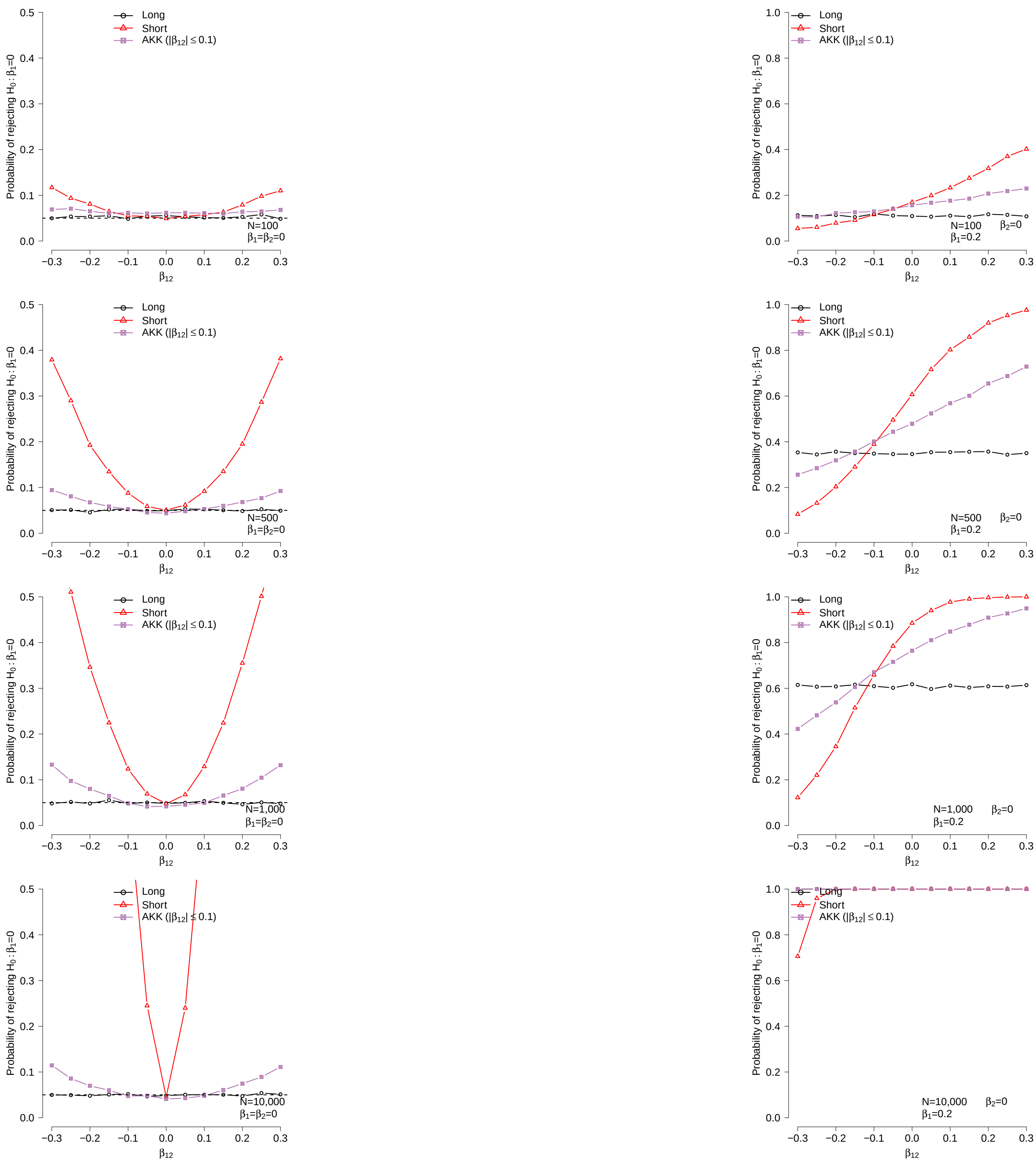


Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$*. In each figure, dashed lines show the power for the long model, while solid lines show power for Elliott et al. (2015)'s nearly optimal test.*

A.6.4 Restrictions on the magnitude of β_{12}: Armstrong et al. (2019)

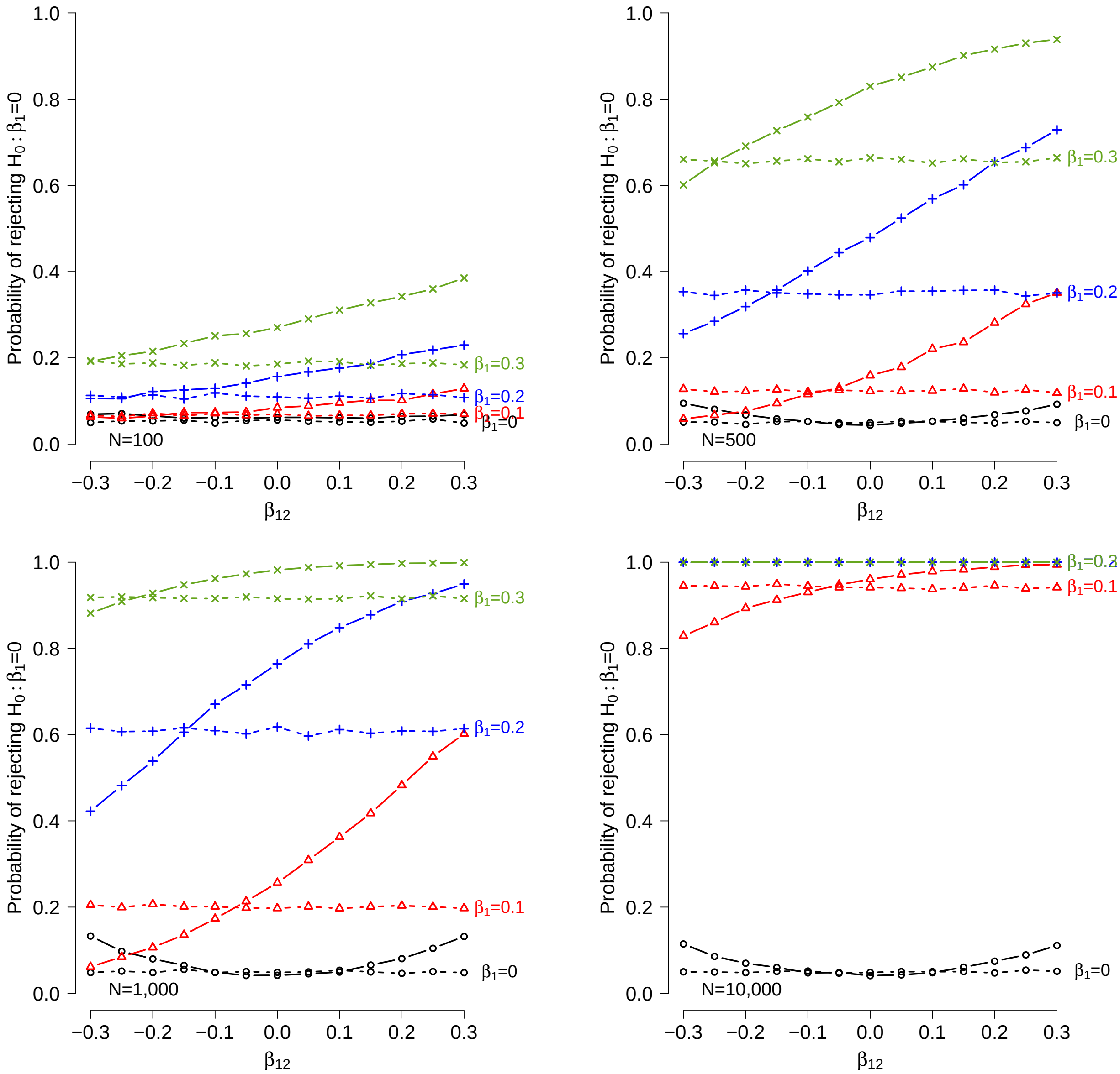
Figure A.6: Armstrong et al. (2019)'s approach: Size and power

(a) Size (b) Power



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

Figure A.7: Long model and Armstrong et al. (2019)'s approach: Power curves



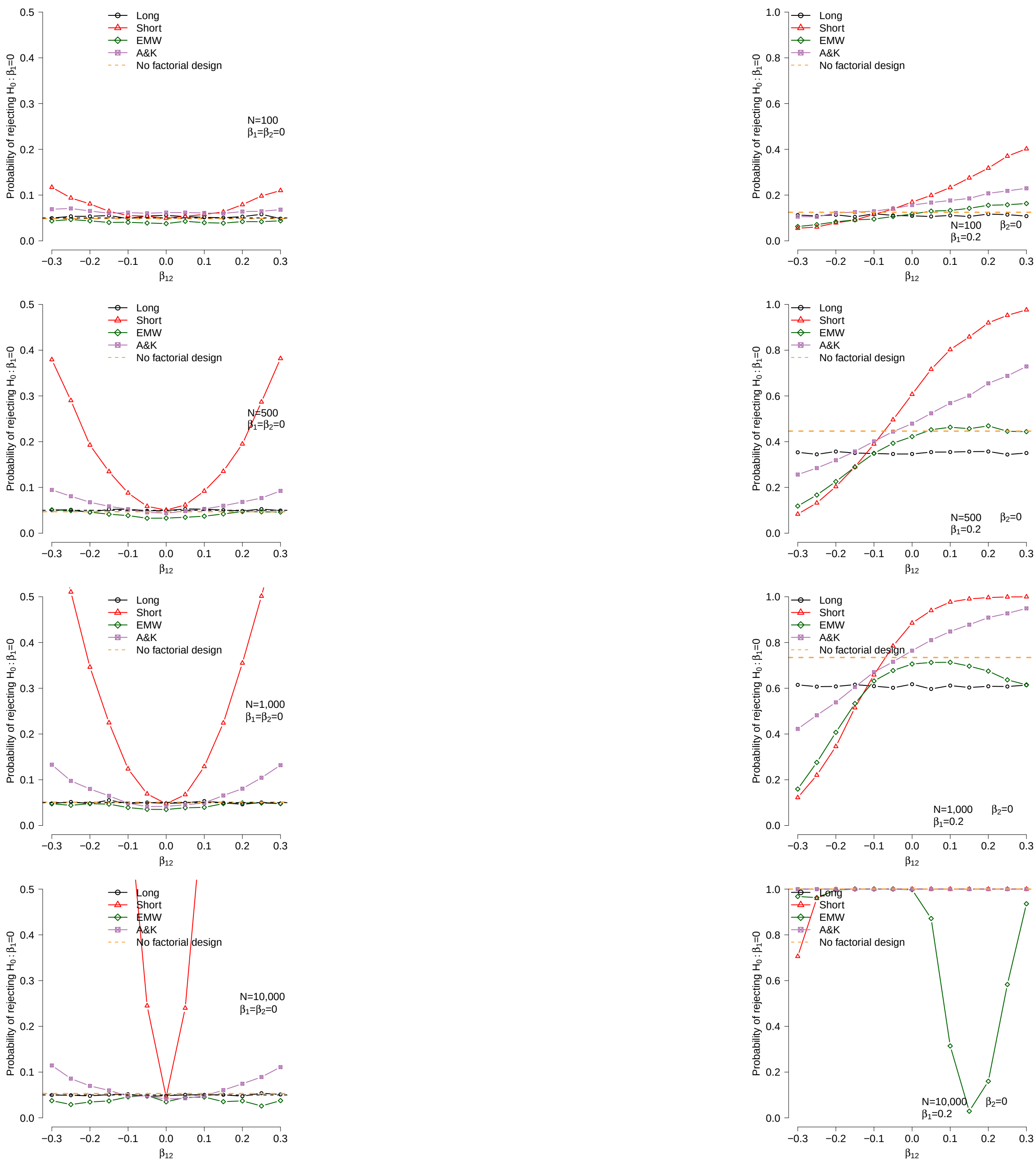
Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. *In each figure, dashed lines show the power for the long model, while solid lines show power for Armstrong et al. (2019)'s approach based on restrictions on the magnitude of* β_{12}.

A.6.5 Leaving the interaction cell empty

Figure A.8: No factorial design: Size and power

(a) Size

(b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. We split the sample size in the interaction cell (N_4) equally distributed the other cells.