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

Karthik Muralidharan† Mauricio Romero‡ Kaspar Wüthrich§

September 13, 2019

Abstract

Cross-cutting or factorial designs are widely used in field experiments. Standard t-tests using the fully-saturated “long” model provide valid inference on the main treatment effects and all interactions. However, t-tests using a “short” model (without interactions) yield greater power for inference on the main treatment effects if the interactions are zero. We show that the assumption of zero interactions is problematic and leads to a significant increase in incorrect inference regarding the main treatment effects relative to a “business as usual” counterfactual. Further, we show that pre-testing the interactions and ignoring them if they are not significant also leads to incorrect inference (due to the implied model selection). We examine econometric approaches to improve power relative to the long model while controlling size for all values of the interaction. Modest “local” power improvements are possible, but come at the cost of lower power for most values of the interaction. For the design of new experiments, an alternative is to leave the interaction cells empty. This design-based approach yields global power improvements while controlling size and we recommend it for policy experiments where a “business as usual” counterfactual is especially important.

Keywords: Randomized Controlled Trial; Factorial designs; Cross-cut designs; Field experiments

JEL Codes: C12, C18, C90, C93

*We are grateful to Isaiah Andrews, Tim Armstrong, Arun Chandrasekhar, Clement de Chaisemartin, Stefano DellaVigna, Esther Duflo, Pascaline Dupas, Graham Elliott, Markus Goldstein, Hiroaki Kaido, Michal Kolesar, Soonwoo Kwon, Craig McIntosh, Adam McCloskey, 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.

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

Cross-cutting or factorial designs are widely used in field experiments to study the effects of multiple treatments in a cost-effective way. However, unbiased estimation and correct inference of the main treatment effects in such experiments depend crucially on the assumption that the interaction between programs is negligible. As Kremer (2003) puts it: “Conducting a series of evaluations in the same area allows substantial cost savings...Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically...This tactic can be problematic, however, if there are significant interactions between programs”.

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

These gains in power, however, come at the cost of an increased likelihood of incorrect inference 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 term. Doing so has non-trivial implications for inference regarding the main treatment effects. The median change in the point estimates of the main treatment effects is 96%; about 26% of estimates change sign; 23% of the estimates switch significance at the 10% level, and 20% do so at the 5% level. Furthermore, 53% (29 out of 55) of estimates reported to be significant at the 5% level are no longer so after including interactions.

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

¹The 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.
the short model if they do not reject that the interaction is zero. However, the distribution of the estimators obtained from this data-dependent model selection procedure are complicated and highly non-normal, making the usual $t$-statistics misleading (Leeb & Pötscher, 2005, 2006, 2008). Further, cross-cutting experiments are rarely adequately powered to detect significant interactions.\footnote{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, 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 inference 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 being conditional on the distribution of the other treatment arms in the experiment. Indeed, the issue that we highlight would not be a concern if we redefined the parameter of interest to be not the main treatment effect, but the composite treatment effect that includes a weighted-average of the interactions with other treatments. However, in practice, papers presenting experimental results from factorial designs, rarely mention this important caveat.\footnote{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 treatment effects should be interpreted this way (or against a counterfactual that was not “business as usual” but one that also had the other treatments in the same experiment).}

This status quo is problematic for at least three reasons. First, ignoring interactions affects the internal validity of experimental estimates. If the interventions studied are new, the other programs may not even exist in the study population. Even if they do, there is no reason to believe that the distributions in the population mirror those in the experiment. Thus, to the extent that estimation and inference of treatment effects in a study population depends on what other interventions are being studied in the same experiment, ignoring interactions is a threat to internal validity.\footnote{Even if we focus only on policy experiments and exclude mechanism experiments (which have more treatments and interactions on average), the rate of false rejection is still 32% (6 out of 19 estimates that are significant at the 5% level without interactions are no longer so after including all interactions).} Policy discussions typically assume a “business as usual” counterfactual, and a treatment effect that is conditional on the nature and distribution of the other treatments in the same experiment will not usually represent a comparison with the policy-relevant counterfactual.\footnote{Some readers may classify the issue we identify as a problem of external rather than internal validity, where unobserved contextual covariates (including other treatment arms) will vary across settings. 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, researchers who use cross-cutting designs often use the two-step procedure described above, and present results from the short model if they do not reject that the interaction is zero.}
Second, interactions are quantitatively important and typically not second-order. We find that the median absolute magnitude of interactions is $0.066\sigma$ and the median absolute value of interactions relative to the main treatment effects is 0.37. The view that interactions are second-order may have been influenced in part by the lack of evidence of significant interactions in most experiments to date. However, this is at least partly because very few experiments are adequately powered to detect interactions. Thus, “absence of evidence” of significant interactions may be getting erroneously interpreted as “evidence of absence.” There is now both experimental (Mbiti et al., 2019; Duflo et al., 2015a) and non-experimental (Kerwin & Thornton, 2017; Gilligan et al., 2018) evidence that interactions matter. Indeed, there is a long tradition in development economics that has highlighted the importance of complementarities across policies/programs in alleviating poverty traps (Ray, 1998; Banerjee & Duflo, 2005), which suggests that assuming away interactions in empirical work may be a mistake.

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 Blair et al. (2019). However, policy experiments are typically used based on whether the intervention had a “significant” effect, both because of publication bias and because of how evidence is aggregated. First, publication-bias towards significant findings is well documented (e.g., I. Andrews & Kasy, 2018; Christensen & Miguel, 2018; Franco et al., 2014). Second, 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 summarized and translated into policy.

Given the issues with interpretation of estimates from the short model, we recommend that all cross-cutting experiments report the results from the long model. Further, the justification for the short model should not be that the interactions were not significant in the long model (because of the model selection issue discussed above). Rather, if researchers would like to present and discuss results from the short model, they should clearly indicate that the presented treatment effects should be interpreted as being con-

---

6The median value of the interaction term across studies is in fact close to zero. The problem is that the median absolute value of the interaction term is not zero, which results in a non-trivial rate of false rejection of the null hypothesis in any given study.
ditional on the distribution of other treatments in the same experiment (and commit to the estimand of interest in a pre-analysis plan). This will ensure transparency in interpretation of the main results as including interactions, and enable readers to assess the extent to which the other treatments may or may not be typical background factors which can be ignored. Further, one can always use the long model to test hypotheses about coefficients from the short model as well as for any policy-relevant allocation of the other treatments and not just the one in the experiment.

The discussions in Kremer (2003) and Duflo et al. (2007) suggest that an important motivation for cross-cutting designs is the belief that interactions are “small” relative to the main treatment effects of interest. We therefore consider if it may be possible to design tests for the main treatment effects relative to a “business as usual” counterfactual that improve power relative to the \( t \)-test based on the long model while maintaining size control for relevant values of the interaction effect.

The two-sided \( t \)-test based on the long regression model is the uniformly most powerful unbiased test (e.g., van der Vaart, 1998). Hence, any procedure that is more powerful than the \( t \)-test for some values of the interactions, must underperform somewhere else. Thus, to achieve higher power, one has to make a choice about where to direct power based on some form of prior knowledge, and pay a price in terms of low power when this prior knowledge is incorrect. Even in the best case, the theoretical scope for power improvements is limited if one insists on size control for all values of the interaction.

We discuss two econometric approaches for improving power. 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 (e.g., near an interaction of zero), while controlling size for all values of the interaction. Our simulation evidence suggests that this approach comes close to achieving the maximal possible power gains relative to the \( t \)-test based on the long model near the likely values of the interaction, but exhibits substantially lower power farther away from this value. Our second approach is to construct confidence intervals for the main treatment effects under prior knowledge on the magnitude of the interaction based on the approach of Armstrong & Kolesar (2018) and Armstrong et al. (2019). We confirm in simulations that, when the prior knowledge is correct, this approach controls size, and also 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. Since the problem

\footnote{For the corresponding one-sided testing problem the one-sided \( t \)-test is uniformly most powerful among all tests that are valid for all values of the interactions. 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 for a 5% test.}
we identify is mainly a result of not knowing the value of the interaction *ex ante*, this
approach may be of limited use in practice.

Our recommendation for the analysis of completed experiments is to 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. For 2×2 factorial designs, if there is
strong prior knowledge about the interaction effect, the nearly optimal test may consti-
tute a suitable alternative and yield modest local power improvements. However, the
nearly optimal test is computationally prohibitive for designs with multiple interaction
cells. Therefore, it cannot be recommended for complicated factorial designs, where the
only viable option may be to report the results from the long model, or to suitably caveat
the interpretation of results from the short model as discussed above.

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

Our most important contribution is to the literature on the design of field experi-
ments. 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 exist-
ing 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

---

8Our code to implement this procedure is available at: https://mtromero.shinyapps.io/elliott/.
9This design will yield power gains relative to running two separate experiments, because the control
group is used twice. But it avoids the problem of interactions noted above. An example of such a design
is provided by Muralidharan & Sundararaman (2011) who study the impact of four different interventions
in one experiment with one common control group, but no cross-cutting treatment arms.
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 $2 \times 2$ 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: $2 \times 2$ factorial design

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_2$</td>
<td>No</td>
<td>$N_1$</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>$N_3$</td>
</tr>
</tbody>
</table>

*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

$$Y = Y_{0,0} \cdot 1_{\{T_1=0,T_2=0\}} + Y_{1,0} \cdot 1_{\{T_1=1,T_2=0\}} + Y_{0,1} \cdot 1_{\{T_1=0,T_2=1\}} + Y_{1,1} \cdot 1_{\{T_1=1,T_2=1\}}, \quad (1)$$
where $I_{\{A\}}$ 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):

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

We refer to $E(Y_{1,0} - Y_{0,0})$ and $E(Y_{0,1} - Y_{0,0})$ 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

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

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

- $E(Y_{1,0} - Y_{0,0}) = E(Y | T_1 = 1, T_2 = 0) - E(Y | T_1 = 0, T_2 = 0)$
- $E(Y_{0,1} - Y_{0,0}) = E(Y | T_1 = 0, T_2 = 1) - E(Y | T_1 = 0, T_2 = 0)$
- $E(Y_{1,1} - Y_{0,1}) = E(Y | T_1 = 1, T_2 = 1) - E(Y | T_1 = 0, T_2 = 1)$
- $E(Y_{1,1} - Y_{1,0}) = E(Y | T_1 = 1, T_2 = 1) - E(Y | T_1 = 1, T_2 = 0)$
- $E(Y_{1,1} - Y_{0,0}) = E(Y | T_1 = 1, T_2 = 1) - E(Y | T_1 = 0, T_2 = 0)$

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:

- $Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 T_2 + \epsilon$, (long model) (3)
- $Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \epsilon^s$, (short model) (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. The coefficients in the long regression model correspond to the main effects of $T_1$ and $T_2$ against a “business as usual” counterfactual and the interaction effect:

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

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

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

By contrast, the regression coefficients in the short model are

$$\beta_1^s = E(Y_{1,1} - Y_{0,1}) P(T_2 = 1) + E(Y_{1,0} - Y_{0,0}) P(T_2 = 0)$$  \hspace{1cm} (8)

$$= E(Y_{1,0} - Y_{0,0}) + E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}) P(T_2 = 1)$$  \hspace{1cm} (9)

$$= \beta_1 + \beta_{12} P(T_2 = 1)$$

and

$$\beta_2^s = E(Y_{1,1} - Y_{1,0}) P(T_1 = 1) + E(Y_{0,1} - Y_{0,0}) P(T_1 = 0)$$  \hspace{1cm} (10)

$$= E(Y_{0,1} - Y_{0,0}) + E(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}) P(T_1 = 1)$$  \hspace{1cm} (11)

$$= \beta_2 + \beta_{12} P(T_1 = 1)$$

Equation (8) shows that $\beta_1^s$ 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, $\beta_2^s$ 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 $\beta_2^s$.

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 weighted average of

---

10The population regression coefficient $\beta$ in the model $Y = X'\beta + \varepsilon$ is the solution to the population least squares problem and is given by $\beta = E(XX')^{-1} E(XY)$.
ATEs relative to different counterfactuals. This composite effect depends on the nature and distribution of the other treatments in the same experiment and is typically neither of primary academic interest nor policy-relevant. Consistent with this view, none of the 19 experimental papers that we re-analyze motivate their experiment as being about estimating this composite effect.

2.3 Long or short: What do we care about?

Section 2.1 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. 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 (or nonparametric) regressions and to make inference on the average treatment effects (e.g., Angrist & Krueger, 1999).

However, experimental treatments are fundamentally different in nature from standard covariates. They are determined by the experimenter based on power and other research-related considerations 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.

Focusing on the short model is particularly problematic in experiments that evaluate policies or programs. Since policy discussions typically assume a “business as usual” counterfactual, a treatment effect that depends on the nature and distribution of the other treatments that just happened to be studied in the same experiment is not informative and can result in misleading policy advice. Moreover, experimental evaluations typically inform policy through systematic reviews and meta-analyses of individual studies. These reviews often simply count the number of studies where an intervention has been found to be effective at conventional significance levels. Thus, the sensitivity of the significance of point estimates to the inclusion/exclusion of interaction terms and the extent of interaction effects with other treatments in the same study, is likely to have non-trivial implications for how evidence is summarized and translated into policy.

There are settings where it may be fine to focus on the short model. For example, in experiments that focus on understanding mechanisms (such as resume audit studies), treat-
ment 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 the 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: \( H_0 : \beta_{s1} = \beta_1 + \beta_{12}P(T_2 = 1) = 0 \). Which test is more powerful depends on the relative magnitude of the four experimental cells.\(^{11}\) 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, \( \beta_{s1} \) is rarely of interest in its own right, and even if it is, the long model allows estimation and inference on \( \beta_{s1} \) 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 \( \beta_1 \). The analysis for \( \beta_2 \) is symmetric and omitted.

Under random assignment and standard regularity conditions, the OLS estimator of \( \beta_1 \) based on the long regression model, \( \hat{\beta}_1 \), is consistent:

\[
\hat{\beta}_1 \xrightarrow{p} \beta_1 = E(Y_{1,0} - Y_{0,0})
\]

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

\[
\hat{\beta}_{s1} \xrightarrow{p} \beta_{s1} = \beta_1 + \beta_{12}P(T_2 = 1).
\]

Unless the true interaction effect is zero (i.e., \( \beta_{12} = 0 \)), \( \hat{\beta}_{s1} \) 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

\(^{11}\)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
\[
Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} T_{2i} + \epsilon_i,
\]
where \( \epsilon_i \sim N(0, \sigma^2) \) and independent of \((T_{1i}, T_{2i})\) and \(\sigma^2\) 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 \((T_{11}, \ldots, T_{1N}, T_{21}, \ldots, T_{2N})\),
\[
\hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1)) \quad \text{and} \quad \hat{\beta}_{1i} \sim N(\beta_{1i}, \text{Var}(\hat{\beta}_{1i})),
\]
where
\[
\text{Var}(\hat{\beta}_{1i}) \leq \text{Var}(\hat{\beta}_1).
\]
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.5 gives explicit formulas of \(\text{Var}(\hat{\beta}_1)\) and \(\text{Var}(\hat{\beta}_1)\) 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.\(^{12}\) 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.\(^{13}\) Only

---


\(^{13}\)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>
<thead>
<tr>
<th></th>
<th>AER</th>
<th>ECMA</th>
<th>JPE</th>
<th>QJE</th>
<th>ReStud</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field experiments</td>
<td>43</td>
<td>9</td>
<td>14</td>
<td>45</td>
<td>13</td>
<td>124</td>
</tr>
<tr>
<td>With factorial designs</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>Interactions included</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Interactions not included</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>19</td>
</tr>
</tbody>
</table>

### 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 $\beta_2$ is symmetric and omitted. We consider a $2 \times 2$ 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

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

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 $0.2\sigma$ 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 individual preference parameters (e.g., Andersen et al. (2008), Gneezy et al. (2009), and Fisman et al. (2008)).
the short regression.\textsuperscript{14} We use Monte Carlo simulations to assess the rejection rates of different procedures under the null (size) and the alternative hypothesis (power).

Figure 1 shows how bias, size and power vary across different values of $\beta_{12}$ in both the long and the short model. As expected, the long model exhibits no bias and correct size for all values of $\beta_{12}$, while the short model has a bias and does not achieve size control when $\beta_{12} \neq 0$. The trade-off is that for $\beta_{12} = 0$, the short model controls size and exhibits higher power than the long model. When $\beta_{12} \neq 0$, the power of the $t$-test based on the short model depends on $\beta_{12}$ 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 $\beta_{12} = 0$.

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 not include all interaction terms. Of these 19, 4 papers did not have publicly-available

\textsuperscript{14}The minimum detectable effect for the long model with power 90\% and size 5\% is $0.29\sigma$. 

Note: Simulations are based on sample size $N$, normal iid errors, and 10,000 repetitions. The size for figures 1b and 1c is $\alpha = 0.05$. 

\textsuperscript{13}
replication data.\footnote{Online Appendix B.1 (in \url{http://mauricio-romero.com/pdfs/papers/Appendix_crosscuts.pdf}) describes the experimental design of each of the 27 papers.}

We downloaded the publicly-available data files and replicated the main results in each of the remaining 15 papers. We then compared the original treatment effects (estimated without the interaction terms) with those estimated including the interaction terms. In other words, we compare estimates based on 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 interaction for these papers is $0\sigma$, the median absolute value of the interaction is $0.066\sigma$. The median absolute value of interactions relative to the main treatment effects is 0.37. 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 interactions being quantitatively important, most experiments will rarely reject the null hypothesis that they are zero (Figure 2 also 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.\footnote{Among the papers that originally included the interaction, 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0% are significant at the 1% level.} 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.

However, Figure 2 highlights that this is mainly because experiments with factorial designs are rarely powered to detect meaningful interactions. This is not surprising, since sample size requirements for detecting interaction effects are onerous. 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.

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 increase 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 $0\sigma$ (dashed vertical line), the median absolute value of the interaction is $0.066\sigma$ (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 change when using the long and the short model. About 48% of treatment estimates that were significant at the 10% level when using the short model are no longer significant when the long model is estimated. 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 when the long regression is estimated (6%, 5%, and 1%, at the 10%, 5%, and 1% levels respectively).

Similar results are seen when we restrict our re-analysis to the ten most cited papers with factorial designs that do not include the interaction terms. Corresponding figures and tables are presented in Appendix A.1.2 (Figure A.2 and Table A.2).

Finally, following the discussion in Section 2.3, we also distinguish between “policy” and “mechanism” 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).
Note: This figure shows how the main treatment estimates change across 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 values 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 0.015σ, the median main treatment estimate from the long model is 0.017σ, the average absolute difference between the treatment estimates of the short and the long model is 0.051σ, 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 from the long and the short regression

<table>
<thead>
<tr>
<th></th>
<th>Panel A: Significance at the 10% level</th>
<th>Panel B: Significance at the 5% level</th>
<th>Panel C: Significance at the 1% level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without interaction</td>
<td>Without interaction</td>
<td>Without interaction</td>
</tr>
<tr>
<td></td>
<td>With interaction</td>
<td>Not significant</td>
<td>Significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>95</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Significant</td>
<td>6</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>101</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

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 estimation and inference of 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). Thus, in practice, 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 $\beta_{12}$ 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 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 estimation of 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.
The main takeaway from Figure 4 is that model selection leads to incorrect inferences and false positives. Thus, researchers should always avoid it.

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 \textit{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 \textit{t}-tests based on the long model, while maintaining size control for relevant values of the interactions. We consider $2 \times 2$ 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 $\beta_1$ and partial out $T_2$, keeping the partialling-out implicit. The analysis for $\beta_2$ is symmetric and omitted. Defining $T_{12} \equiv T_1 T_2$, the
regression model of interest is

\[ Y = \beta_1 T_1 + \beta_{12} T_{12} + \epsilon. \]  

(12)

Our goal is to test hypotheses about the main effect \( \beta_1 \).

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). 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 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).

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 based on Armstrong & Kolesar (2018) and Armstrong

---

17 We omit the intercept because all variables have mean zero after partialling-out \( T_2 \).
18 A test is unbiased if its power is larger than its size.
19 Moreover, the results in Joshi (1969) imply the usual two-sided confidence interval based on the long regression model achieves minimax expected length; see also Armstrong & Kolesar (2019).
et al. (2019). 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.7 explores two additional econometric approaches based on work by Imbens & Manski (2004); Stoye (2009) and McCloskey (2017).

5.2 Nearly optimal tests targeting power towards a likely value \( \beta_{12} = \bar{\beta}_{12} \)

Consider a scenario where a particular value \( \beta_{12} = \bar{\beta}_{12} \) seems a priori likely and suppose that we want to find a test that controls size and is as powerful as possible when \( \beta_{12} = \bar{\beta}_{12} \). For concreteness, we focus on the case where \( \bar{\beta}_{12} = 0 \) and consider the following testing problem

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

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 \( \hat{t}_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \), \( \hat{t}_{12} = \frac{\hat{\beta}_{12}}{SE(\hat{\beta}_{12})} \), \( t_1 = \frac{\beta_1}{SE(\beta_1)} \), \( t_{12} = \frac{\beta_{12}}{SE(\beta_{12})} \), and \( \rho = \text{Cov}(t_1, t_{12}) \). 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 \( \text{Cov}(t_1, t_{12}) \) 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 “\( H_0: \text{the density of } \hat{t} \text{ is } f_t, t_1 = 0, t_{12} \in \mathbb{R} \)” against the simple alter-
native “$H_{1,F}$: 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), $\Lambda^{LF}$, with the property that the size $\alpha$ Neyman-Pearson test of the simple hypothesis $H_{0,\Lambda^{LF}}$ against $H_{1,F}$ also yields a size $\alpha$ 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 $\Lambda^{LF}$, Elliott et al. (2015) suggest to instead focus on an approximate LFD, $\Lambda^{ALF}$, 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 $\Lambda^{ALF}$.

Figure 5: Elliott et al. (2015)’s nearly optimal test controls for size and has power gains over running the full model for “intermediate” values of $\beta_{12}$

(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.\textsuperscript{20} The test controls size for all values of $\beta_{12}$ and, by construction, is nearly optimal at $\beta_{12} = 0$. A comparison with the $t$-test based on the long model shows that the nearly optimal test is more powerful when $\beta_{12}$ is close to zero. The\textsuperscript{20}

\textsuperscript{20}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.
nearly optimal test comes close to achieving the maximal possible power. However, as expected given the discussion in Section 5.1, these power gains come at a cost: for certain values of $\beta_{12}$, the power can be much lower than the power of the $t$-test based on the long model. Appendix A.8.3 provides a more comprehensive assessment of the performance of the nearly optimal tests by plotting power curves for different values of $\beta_1$.

5.3 Inference under a priori restrictions on the magnitude of $\beta_{12}$

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 $\beta_{12}$ 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 $\beta_{12}$.

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

Assumption 1 restricts the parameter space for $\beta_{12}$ and implies that

$$\beta_{12} \in \{b_{12} : |b_{12}| \leq C\} \equiv B_{12}.$$  

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

$$Y = \beta_1 T_1 + \beta_{12} T_{12} + \epsilon$$  \hspace{1cm} (16)

and assume that $\epsilon \sim N(0, \sigma^2 I_N)$ and that $\sigma^2$ is known. The implementation with heteroskedastic and non-Gaussian errors is discussed in Appendix A.6. An affine estimator of $\beta_1$ can be written as $\hat{\beta}_1 = a + b'Y$, for some $a$ and $b$ that can depend on $X \equiv (T_1, T_{12})$. For example, for the long OLS regression model, $a = 0$ and $b$ is the first row of $(X'X)^{-1}X'$.

---

21 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.

22 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.
Define the “worst case” biases as

\[
\text{Bias}(\hat{\beta}_1) = \sup_{\beta_1 \in \mathbb{R}, \beta_{12} \in B_{12}} E_{(\beta_1, \beta_{12})} (\hat{\beta}_1 - \beta_1),
\]

\[
\text{Bias}(\hat{\beta}_1) = \inf_{\beta_1 \in \mathbb{R}, \beta_{12} \in B_{12}} E_{(\beta_1, \beta_{12})} (\hat{\beta}_1 - \beta_1),
\]

where \(E_{(\beta_1, \beta_{12})}\) denotes the expectation under the distribution generated by model (16) with \((\beta_1, \beta_{12})\). 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 \((\beta_1, \beta_{12})\) varies over \(\mathbb{R} \times B_{12}\), the \(t\)-ratio, \(t \equiv \frac{(\hat{\beta}_1 - \beta_1)}{SE(\hat{\beta}_1)}\), is normally distributed with variance one and mean varying from \(\frac{\text{Bias}(\hat{\beta}_1)}{SE(\hat{\beta}_1)}\) to \(\frac{\text{Bias}(\hat{\beta}_1)}{SE(\hat{\beta}_1)}\). To construct a two-sided confidence interval, note that testing \(H_0: \beta_1 = \beta_0\) based on a \(t\)-statistic with critical value \(cv_\alpha\left(\max\{|\text{Bias}(\hat{\beta}_1)|,|\text{Bias}(\hat{\beta}_1)|\}\right) SE(\hat{\beta}_1)\) yields a level \(\alpha\) 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:

\[
\hat{\beta}_1 \pm cv_\alpha\left(\max\{|\text{Bias}(\hat{\beta}_1)|,|\text{Bias}(\hat{\beta}_1)|\}\right) SE(\hat{\beta}_1) \quad (17)
\]

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.6.

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 \(B_{12} = [-0.1, 0.1]\). 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 \(|\beta_{12}| > C\). Appendix A.8.4 presents the corresponding power curves for different values of \(\beta_1\).
Figure 6: Restrictions on the magnitude of $\beta_{12}$ yield power gains if they are correct but lead to incorrect inferences if they are not correct.

(a) Size

(b) Power

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

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 natural alternative is to leave the “interaction cell” empty (i.e., to set $N_4 = 0$) and to re-assign those subjects to the other cells such that $P(T_1 = 1, T_2 = 1) = 0$. We show in Appendix A.4.1 that if $T_1$ and $T_2$ are randomly assigned,

$$
\beta_{1s} = E(Y_{1,0} - Y_{0,0}),
\beta_{2s} = E(Y_{0,1} - Y_{0,0}).
$$

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

Consider next the power implications of leaving the interaction cell empty. Specifically, note that $P(T_1 = 1, T_2 = 1) = 0$ implies that $T_1$ and $T_2$ are not independent if $P(T_1 = 1)$ and $P(T_2 = 1)$ are both non-zero.
consider an experiment where the researcher cares equally about power to detect an effect for $T_1$ and $T_2$, and thus assigns the same sample size to both treatments:

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T_2$</th>
<th>No</th>
<th>NC</th>
<th>NT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>N</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

In what follows, we focus on $\beta_{s1}$. The analysis for $\beta_{s2}$ is symmetric and omitted. Let $\hat{\beta}_{1nf}$ denote the OLS estimator based on the short model, leaving the interaction cell empty. Under the assumptions of Section A.5.1, the variance of $\hat{\beta}_{1nf}$ is given by

$$Var(\hat{\beta}_{1nf}) = \sigma^2 \frac{N - NT}{(N - 2NT)NT}.$$ 

$Var(\hat{\beta}_{1nf})$ is minimized when $NT = N \left(2 - \sqrt{2}\right)$ and we assume that the experiment is designed in this manner.\(^{24}\) A comparison with the variance of the estimator based on the long model, $\hat{\beta}_1$, is

$$Var(\hat{\beta}_{1nf}) \leq Var(\hat{\beta}_1).$$\(^{25}\) Thus, by the same reasoning as in Section 2.4, leaving the interaction cell empty leads to power improvements for testing hypotheses about $\beta_1$ relative to the long regression model.

Figure 7 presents the results based on our running example. As expected, leaving the interaction cell empty yields tests that control size for all values of the interaction. Moreover, among the approaches that achieve size control for all values of $\beta_{12}$ (the long model and the nearly optimal test), leaving the interaction cell empty yields the highest power. Thus, if one is not interested in interaction effects, we recommend avoiding factorial designs and leaving the interaction cell empty as done for instance in the design employed by Muralidharan & Sundararaman (2011).

\(^{24}\)This exact sample split is impossible in any application since $N \left(2 - \sqrt{2}\right)$ is not an integer. In our simulations we therefore use $NT = 0.29N$ and $NC = 0.42N$.

\(^{25}\)Here, we implicitly assume that the goal is to have equal power to detect an effect for $T_1$ and $T_2$. 

27
Figure 7: Leaving the interaction cell empty increases power for most values of $\beta_{12}$ relative to other alternatives

(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 to construct optimal confidence intervals under prior knowledge.

7 Factorial designs with more than two treatments

So far, our theoretical discussion has focused on $2 \times 2$ 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.

The $t$-tests based on the short model are more powerful than the corresponding $t$-tests based on the long model when the interactions are zero, but may suffer from size distortions when they are not. Conceptually, both econometric approaches discussed in Section 5 can be extended beyond $2 \times 2$ 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.\textsuperscript{26}

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 main effects relative to a “business as usual” counterfactual are important, we recommend leaving the interaction cells empty at the design stage, which will yield 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. This is the approach we recommend for policy experiments where a “business as usual” counterfactual is especially important.

In some cases (such as resume audit experiments aiming to study the existence of certain mechanisms), the short model may be fine. But the estimand of interest should be committed to in a pre-analysis plan and authors should be explicit in both papers and pre-analysis plans that the estimated effect includes both the main effect and a

\textsuperscript{26}Both approaches discussed in the Appendix A.7 are computationally feasible in more complicated cross-cut designs.
weighted-average of interactions with other treatments in the experiment.

The issue we document in this paper and our recommendations for practice are in part motivated by how evidence is aggregated in practice, which usually reflects a frequentist approach. Experimental evaluations typically inform policy through systematic reviews and meta-analyses of individual studies. These reviews often simply count the number of studies where an intervention has been found to be effective at conventional significance levels. Thus, the sensitivity of the significance of point estimates to the inclusion/exclusion of interaction terms and the extent of interaction effects with other treatments in the same study, is likely to have non-trivial implications for how evidence is summarized and translated into policy.

The conclusions and practical recommendations can be different if 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 in settings where the goal is to make better decisions in the specific setting of the experiment. Cross-cutting experiments also make sense in a setting of iterative high-frequency experiments, where the goal for initial experiments is to explore several treatment dimensions in an efficient way, to generate promising interventions for further testing.

However, policy experiments are expensive and difficult to run iteratively, and are typically used based on whether the intervention had a “significant” effect. This concern also applies to publication itself, where publication-bias towards significant findings is well documented (e.g., I. Andrews & Kasy, 2018; Christensen & Miguel, 2018; Franco et al., 2014). This is the setting where the factorial design is problematic for the reasons documented in this paper.

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

---

27This 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 is less about testing whether any of these factors “matter” (in the sense of being significant) and more to make better informed decisions regarding the optimal use of various inputs. These are cases where a loss-function approach makes a lot of sense and the factorial design is likely to be more efficient than a non-interacted one.
References


# 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 April of 2018)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>Citations</th>
<th>Treatments</th>
<th>Interactions In Design</th>
<th>Interactions Included</th>
<th>Data Available</th>
<th>Policy Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olken (2007)</td>
<td>Monitoring Corruption: Evidence from a Field Experiment in Indonesia</td>
<td>JPE</td>
<td>2007</td>
<td>1529</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Banerjee et al. (2007)</td>
<td>Remedying Education: Evidence from Two Randomized Experiments in India</td>
<td>QJE</td>
<td>2007</td>
<td>1213</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Kleven et al. (2011)</td>
<td>Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark</td>
<td>ECMA</td>
<td>2011</td>
<td>776</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Karlan et al. (2014)</td>
<td>Agricultural Decisions after Relaxing Credit and Risk Constraints</td>
<td>QJE</td>
<td>2014</td>
<td>612</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Bertrand et al. (2010)</td>
<td>What’s Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment</td>
<td>QJE</td>
<td>2010</td>
<td>522</td>
<td>14</td>
<td>85</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>Citations</th>
<th>Treatments</th>
<th>Interactions In Design</th>
<th>Interactions Included</th>
<th>Data Available</th>
<th>Policy Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karlan &amp; List (2007)</td>
<td>Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment</td>
<td>AER</td>
<td>2007</td>
<td>506</td>
<td>7</td>
<td>28</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Thornton (2008)</td>
<td>The Demand for, and Impact of, Learning HIV Status</td>
<td>AER</td>
<td>2008</td>
<td>453</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Haushofer &amp; Shapiro</td>
<td>The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya</td>
<td>QJE</td>
<td>2016</td>
<td>393</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Alatas et al. (2012)</td>
<td>Targeting the Poor: Evidence from a Field Experiment in Indonesia</td>
<td>AER</td>
<td>2012</td>
<td>330</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Karlan &amp; Zinman (2008)</td>
<td>Credit Elasticities in Less-Developed Economies: Implications for Microfinance</td>
<td>AER</td>
<td>2008</td>
<td>311</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Duflo et al. (2015a)</td>
<td>Education, HIV, and Early Fertility: Experimental Evidence from Kenya</td>
<td>AER</td>
<td>2015</td>
<td>282</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Andreoni et al. (2017)</td>
<td>Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving</td>
<td>JPE</td>
<td>2017</td>
<td>270</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>Citations</th>
<th>Treatments</th>
<th>Interactions In Design</th>
<th>Interactions Included</th>
<th>Data Available</th>
<th>Policy Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eriksson &amp; Rooth (2014)</strong></td>
<td>Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment</td>
<td>AER</td>
<td>2014</td>
<td>238</td>
<td>34</td>
<td>71680</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Allcott &amp; Taubinsky (2015)</strong></td>
<td>Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market</td>
<td>AER</td>
<td>2015</td>
<td>237</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Flory et al. (2014)</strong></td>
<td>Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions</td>
<td>ReStud</td>
<td>2015</td>
<td>204</td>
<td>10</td>
<td>24</td>
<td>12</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Brown et al. (2010)</strong></td>
<td>Shrouded Attributes and Information Suppression: Evidence from the Field</td>
<td>QJE</td>
<td>2010</td>
<td>189</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>DellaVigna et al. (2016)</strong></td>
<td>Voting to Tell Others Contract Structure, Risk-Sharing, and Investment Choice</td>
<td>ReStud</td>
<td>2017</td>
<td>169</td>
<td>4</td>
<td>15</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Fischer (2013)</strong></td>
<td></td>
<td>ECMA</td>
<td>2013</td>
<td>162</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Kaur et al. (2015)</strong></td>
<td>Self-Control at Work</td>
<td>JPE</td>
<td>2015</td>
<td>154</td>
<td>8</td>
<td>16</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Cohen et al. (2015)</strong></td>
<td>Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial</td>
<td>AER</td>
<td>2015</td>
<td>151</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>Citations</th>
<th>Treatments</th>
<th>Interactions In Design</th>
<th>Interactions Included</th>
<th>Data Available</th>
<th>Policy Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blattman et al. (2017)</td>
<td>Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia</td>
<td>AER</td>
<td>2017</td>
<td>135</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Khan et al. (2015)</td>
<td>Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors</td>
<td>QJE</td>
<td>2016</td>
<td>133</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Balafoutas et al. (2013)</td>
<td>What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods</td>
<td>ReStud</td>
<td>2013</td>
<td>126</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Kendall et al. (2015)</td>
<td>How Do Voters Respond to Information? Evidence from a Randomized Campaign</td>
<td>AER</td>
<td>2015</td>
<td>116</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Pallais &amp; Sands (2016)</td>
<td>Why the Referential Treatment? Evidence from Field Experiments on Referrals</td>
<td>JPE</td>
<td>2016</td>
<td>85</td>
<td>3</td>
<td>12</td>
<td>0</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: This table provides relevant information from the most cited articles with factorial designs. Citation counts are from Google Scholar on July 4th of 2019. Treatments is the number of different treatments in the paper. “Interactions in Design” is the number of interactions in the experimental design. “Interactions Included” is the number of interactions included in the main specification of the paper. Data available refers to whether the data is publicly available or not. Allcott & Taubinsky (2015) has two field experiments. The table refers to the second one. Section B.1.16 provides for more details. Flory et al. (2014) one of the three dimensions of randomization 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 - All

Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.
A.1.2 Ten Most Cited Papers

Figure A.2: Treatment estimates from the long and the short regression

(a) Main treatment estimates

(b) Interaction

Note: Both figures show treatment estimates from the ten most cited papers with factorial designs and publicly available data that do not include the interaction in the original study. Figure A.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 $0.007\sigma$, the median main treatment estimate from the long model is $0.014\sigma$, the average absolute difference between the treatment estimates of the short and the long model is $0.05\sigma$, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 130.7%, and 28.2% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.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 $-0.003\sigma$ (dashed vertical line), the median absolute value of the interactions is $0.051\sigma$ (dashed vertical line), 5.6% of interactions are significant at the 10% level, 2.6% are significant at the 5% level, and 0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.56.
Table A.2: Significance of treatment estimates from the long and the short regression

<table>
<thead>
<tr>
<th></th>
<th>Without interaction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
<td>Significant</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Not significant</td>
<td>49</td>
<td>13</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Significant</td>
<td>6</td>
<td>17</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>30</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

Panel B: Significance at the 5% level

<table>
<thead>
<tr>
<th></th>
<th>Without interaction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
<td>Significant</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Not significant</td>
<td>60</td>
<td>9</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Significant</td>
<td>4</td>
<td>12</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>21</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

Panel C: Significance at the 1% level

<table>
<thead>
<tr>
<th></th>
<th>Without interaction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
<td>Significant</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Not significant</td>
<td>73</td>
<td>3</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Significant</td>
<td>1</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td>11</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

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 - ten most cited

Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.
A.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 0.057σ, the median main treatment estimate from the long model is 0.054σ, the average absolute difference between the treatment estimates of the short and the long model is 0.069σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 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% 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

<table>
<thead>
<tr>
<th>Panel A: Significance at the 10% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>31</td>
</tr>
<tr>
<td>Significant</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Significance at the 5% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>43</td>
</tr>
<tr>
<td>Significant</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Significance at the 1% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>56</td>
</tr>
<tr>
<td>Significant</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
</tr>
</tbody>
</table>

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 - policy evaluations

Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical line is at 1.96. That is, when the t-value is to the right of the vertical line, the interaction is significant at the 5% level.
A.1.4 Studies with all interactions included

Figure A.6: 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 include the interaction in the original study. 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 -0.031σ, the median main treatment estimate from the long model is -0.016σ, the average absolute difference between the treatment estimates of the short and the long model is 0.046σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 36.9%, and 14.5% 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 -0.008σ, 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.52.
Table A.4: Significance of treatment estimates from the long and the short regression

<table>
<thead>
<tr>
<th>Panel A: Significance at the 10% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>61</td>
</tr>
<tr>
<td>Significant</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Significance at the 5% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>68</td>
</tr>
<tr>
<td>Significant</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Significance at the 1% level</th>
<th>Without interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
</tr>
<tr>
<td>Not significant</td>
<td>77</td>
</tr>
<tr>
<td>Significant</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>79</td>
</tr>
</tbody>
</table>

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 - studies with all interactions included

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

A.2 Derivation of expressions for the regression coefficients

A.3 Derivation of the expressions for $\beta_1$ and $\beta_2$

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

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

$$
\beta_2 = E(Y \mid T_1 = 0, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 0),
$$

$$
\beta_{12} = E(Y \mid T_1 = 1, T_2 = 1) - E(Y \mid T_1 = 0, T_2 = 1)
\quad - [E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0)].
$$

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

$$
E(Y \mid T_1 = t_1, T_2 = t_2) = E(Y_{t_1,t_2} \mid T_1 = t_1, T_2 = t_2)
\quad = E(Y_{t_1,t_2}).
$$
Thus, it follows that

\[ \beta_1 = E (Y_{1,0} - Y_{0,0}), \]
\[ \beta_2 = E (Y_{0,1} - Y_{0,0}), \]
\[ \beta_{12} = E (Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}). \]

### A.4 Derivation of the expressions for \( \beta^s_1 \) and \( \beta^s_2 \)

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

\[ \beta^s_1 = E (Y | T_1 = 1) - E (Y | T_1 = 0). \]

Consider

\[ E (Y | T_1 = 1) = E (Y | T_1 = 1, T_2 = 1) P(T_2 = 1 | T_1 = 1) + E (Y | T_1 = 1, T_2 = 0) P(T_2 = 0 | T_1 = 1) = E (Y_{1,1}) P(T_2 = 1) + E (Y_{1,0}) P(T_2 = 0), \]

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

\[ E (Y | T_1 = 0) = E (Y_{0,1}) P(T_2 = 1) + E (Y_{0,0}) P(T_2 = 0) \]

Thus, we have

\[ \beta^s_1 = E (Y | T_1 = 1) - E (Y | T_1 = 0) = E (Y_{1,1} - Y_{0,1}) P(T_2 = 1) + E (Y_{1,0} - Y_{0,0}) P(T_2 = 0). \]

### A.4.1 Derivation of the expressions for \( \beta^s_1 \) and \( \beta^s_2 \) when the interaction cell is empty

Here we derive the expressions for \( \beta^s_1 \) and \( \beta^s_2 \) if we leave the interaction cell empty, i.e., \( P(T_1 = 1, T_2 = 1) = 0 \). To simplify the exposition, we define \( p_1 := P(T_1 = 1), \)
\( p_2 := P(T_2 = 1) \) and \( p_{12} := P(T_1 = 1, T_2 = 1) \). The population regression coefficients \( \beta^s = (\beta^s_0, \beta^s_1, \beta^s_2)' \) is given by

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

where \( X = (1, T_1, T_2)' \). In what follows, we focus on \( \beta^s_1 \); the derivation for \( \beta^s_2 \) is similar. Multiplying out yields the following expressions for \( \beta^s_1 \):

\[
\beta^s_1 = \frac{(p_2 p_{12} - p_1 p_2)E(Y) + p_1(p_2 - p_2^2)E(Y \mid T_1 = 1) + p_2(p_1 p_2 - p_{12})E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2 + 2p_1 p_2 p_{12} - p_{12}^2}.
\]

Using \( p_{12} = 0 \), obtain

\[
\beta^s_1 = \frac{-p_1 p_2 E(Y) + p_1 p_2 (1 - p_2)E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2} \tag{18}
\]

Because \( p_{12} = 0 \), we have that

\[
E(Y) = E(Y \mid T_1 = 1, T_2 = 0) p_1 + E(Y \mid T_1 = 0, T_2 = 0) (1 - p_1 - p_2) + E(Y \mid T_1 = 0, T_2 = 1) p_2.
\]

Combining (18) and (19) and simplifying the yields:

\[
\beta^s_1 = \frac{-p_1 p_2 (E(Y \mid T_1 = 1, T_2 = 0) p_1 + E(Y \mid T_1 = 0, T_2 = 0) (1 - p_1 - p_2))}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2}

- \frac{E(Y \mid T_1 = 0, T_2 = 1) p_2) + p_1 p_2 (1 - p_2) E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{p_1 p_2 (1 - p_1 - p_2)}

\[
= \frac{p_1 p_2 (E(Y \mid T_1 = 1, T_2 = 0) (1 - p_1 - p_2) - E(Y \mid T_1 = 0, T_2 = 0) (1 - p_1 - p_2))}{p_1 p_2 (1 - p_1 - p_2)}

= E(Y \mid T_1 = 1, T_2 = 0) - E(Y \mid T_1 = 0, T_2 = 0)
\]

The result now follows by random assignment of \( T_1 \) and \( T_2 \) and the definition of potential outcomes.
A.5 Variance reductions and power gains based on the short model

A.5.1 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 \( \sigma^2 \) 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 \( \beta_1 \). The analysis for \( \beta_2 \) is symmetric and omitted.

Define \( T_1 \equiv (T_{11}, \ldots, T_{1N})' \) and \( T_2 \equiv (T_{21}, \ldots, T_{2N})' \). If the interaction effect is zero (i.e., \( \beta_{12} = 0 \)), it follows from standard results that conditional on \( (T_1, T_2) \), \( \hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1)) \) and \( \hat{\beta}^z_1 \sim N(\beta_1, \text{Var}(\hat{\beta}^z_1)) \), where

\[
\text{Var}(\hat{\beta}_1) = \sigma^2 \frac{N_1 + N_2}{N_1 N_2} \quad \text{and} \quad \text{Var}(\hat{\beta}^z_1) = \sigma^2 \frac{N_1 N_3 + N_1 N_4 + N_2 N_3 + N_2 N_4}{N_1 N_2 N_3 + N_1 N_2 N_4 + N_1 N_3 N_4 + N_2 N_3 N_4}.
\]

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 \( \hat{\beta}^z_1 \), the \( t \)-test based on the short model exhibits higher finite sample power than the \( t \)-test based on the long model.\(^{28}\)

Let \( \hat{\beta}^z_1 = \hat{\beta}^z_1 / \text{SE}(\hat{\beta}^z_1) \) and \( \hat{\beta}_1 = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1) \), let \( P_{\beta_1} \) denote probabilities under the assumption that \( \beta_1 \) is the true coefficient and let \( c_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) \), where \( \Phi^{-1} \) 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 are given as

\[
P_{\beta_1}(|\hat{\beta}| > c_{1-\alpha/2} | T_1, T_2) = \Phi \left( \frac{\beta_1}{\text{SE}(\hat{\beta}_1)} - c_{1-\alpha/2} \right) + 1 - \Phi \left( \frac{\beta_1}{\text{SE}(\hat{\beta}_1)} + c_{1-\alpha/2} \right),
\]

\(^{28}\)To see this, note that

\[
\text{Var}(\hat{\beta}_1) - \text{Var}(\hat{\beta}^z_1) = \sigma^2 \frac{N_3 N_4 (N_1 + N_2)^2}{N_1 N_2 (N_1 N_2 N_3 + N_1 N_2 N_4 + N_1 N_3 N_4 + N_2 N_3 N_4)} \geq 0.
\]
and

\[ P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2) = \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} - c_{1 - \alpha/2} \right) + 1 - \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} + c_{1 - \alpha/2} \right). \]

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

\[ P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2) \geq P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2). \]

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

\[ \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \mid T_1, T_2 \sim N(0,1) \]

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

\[ P_{\hat{\beta}_1} \left( \hat{\beta} > z \mid T_1, T_2 \right) = P_{\hat{\beta}_1} \left( \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} > z \mid T_1, T_2 \right) \]
\[ = P_{\hat{\beta}_1} \left( \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} > z - \frac{\beta_1}{SE(\hat{\beta}_1)} \mid T_1, T_2 \right) \]
\[ = \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} - z \right). \]

Thus, the power of a two-sided test is

\[ P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2) = P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2) + P_{\hat{\beta}_1} (|\hat{\beta}| < -c_{1 - \alpha/2} \mid T_1, T_2) \]
\[ = \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} - c_{1 - \alpha/2} \right) + 1 - \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} + c_{1 - \alpha/2} \right). \]

Similarly, one can show that

\[ P_{\hat{\beta}_1} (|\hat{\beta}| > c_{1 - \alpha/2} \mid T_1, T_2) = \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} - c_{1 - \alpha/2} \right) + 1 - \Phi \left( \frac{\beta_1}{SE(\hat{\beta}_1)} + c_{1 - \alpha/2} \right). \]

Part (ii): To establish this 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 \( \hat{\beta}_1 \) which satisfies

\[
\tilde{t} \equiv \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta})} \mid T_1, T_2 \sim N(0, 1)
\]

is given by

\[
P_{\hat{\beta}_1} (|\tilde{t}| > c_{1-\alpha/2} \mid T_1, T_2) = \Phi \left( \frac{\beta_1}{SE(\hat{\beta})} - c_{1-\alpha/2} \right) + 1 - \Phi \left( \frac{\beta_1}{SE(\hat{\beta})} + c_{1-\alpha/2} \right).
\]

Consider\(^{29}\)

\[
\frac{\partial P_{\hat{\beta}_1} (|\tilde{t}| > c_{1-\alpha/2} \mid T_1, T_2)}{\partial SE(\hat{\beta})} = \frac{\beta_1}{SE(\hat{\beta})^2} \left[ \phi \left( \frac{\beta_1}{SE(\hat{\beta})} + c_{1-\alpha/2} \right) - \phi \left( \frac{\beta_1}{SE(\hat{\beta})} - c_{1-\alpha/2} \right) \right] \leq 0,
\]

which follows from the shape of the normal distribution.

A.5.2 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 \( 2 \times 2 \) factorial design, the variance of the estimate of \( \beta_1 \) is given by

\[
\text{Var} (\hat{\beta}_1) = \sigma^2 \frac{N_1 + N_2}{N_1 N_2} \quad \text{and} \quad \text{Var} (\hat{\beta}_s) = \sigma^2 \frac{N_1 N_3 + N_1 N_4 + N_2 N_3 + N_2 N_4}{N_1 N_2 N_3 + N_1 N_2 N_4 + N_1 N_3 N_4 + N_2 N_3 N_4}.
\]

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

\[
\text{Var} (\hat{\beta}_1) \equiv \sigma^2 \frac{6}{(1 - \alpha) N} \quad \text{and} \quad \text{Var} (\hat{\beta}_s) \equiv \sigma^2 \frac{6(1 + 2\alpha)}{(1 - \alpha) N(1 + 8\alpha)}.
\]

Figure A.1 shows how the variance changes for different values of \( \alpha \). 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

\(^{29}\)See, for example, Lemma 2 in Carneiro et al. (2017) for a similar argument.
highest power (lowest variance) is achieved when the sample size is allocated equally across cells (i.e., $\alpha = 0.25$). Intuitively, given that we ignore the fact that some individuals get both treatments, at this point the size of the treatment and the control group for $T_1$ is the same.

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

---

A.6 Implementation details for Section 5.3

Recall that under Assumption 1, $\beta_{12} \in \{b_{12} : |b_{12}| \leq C\} \equiv B_{12}$. 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:

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

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

\[
SE(\hat{\beta}_\lambda) \equiv \sqrt{\hat{\sigma}^2 \left\| T_1 - T_{12} \pi \lambda \right\|^2 \left( (T_1 - T_{12} \pi \lambda)' T_1 \right)^2}
\]

\[
Bias(\hat{\beta}_\lambda) \equiv C \frac{(T_1 - T_{12} \pi \lambda)' T_{12} \pi \lambda}{|\pi \lambda| (T_1 - T_{12} \pi \lambda)' T_1}
\]

and \( \pi \lambda \) solves \( \min_{\pi} \| T_1 - \pi T_{12} \|^2 + \lambda |\pi| \). Denote the solution by \( \lambda^* \).

3. Construct the optimal confidence interval:

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

where

\[
\hat{\beta}_{\lambda^*} = \frac{(T_1 - T_{2 \pi \lambda^*})' Y}{(T_1 - T_{2 \pi \lambda^*})' T_1}.
\]

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

### A.7 Additional econometric approaches

In this section, we discuss two additional econometric approaches.

#### A.7.1 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 \( \beta_{12} \). As a consequence, one can search for the largest critical value over the values of \( \beta_{12} \) in the confidence interval rather than over the whole parameter space as in the LF approach. The uncertainty about the nuisance parameter (\( \beta_{12} \)) 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.³¹

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 α = 0.05. For the model selection, the short model is estimated if one fails to reject β₁₂ = 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 β₁₂. 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.³²

³¹We 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.

³²This 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 severely limited.
A.7.2 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 \( \beta := (\beta_0, \beta_1, \beta_2) \) from a population regression of \( Y - \bar{\beta}_{12}T_{12} \) on \( T_1 \) and \( T_2 \). Letting \( X := (1, T_1, T_2)' \), the resulting regression population regression coefficients are given as

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

Assumption 1 implies that \( \beta_{12} \) lies in a compact interval,

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

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

\[
\beta(\beta_{12}) \equiv (XX')^{-1} E(X(Y - \beta_{12}T_{12})) = (XX')^{-1} E(XY) - \beta_{12}E(XX')^{-1} E(XT_{12})
\]

Note that \( (XX')^{-1} E(XT_{12}) \equiv (\gamma_0, \gamma_1, \gamma_2)' \) is the population regression coefficient from a regression of \( T_{12} \) on \( X \). Independence of \( T_1 \) and \( T_2 \) implies that \( \gamma_1 = E(T_{12} | T_1 = 1) - E(T_{12} | T_1 = 0) \) and \( \gamma_2 = E(T_{12} | T_2 = 1) - E(T_{12} | T_2 = 0) \) both of which are positive. Consequently, the identified set for \( \beta_t, t \in \{1, 2\} \), is given by

\[
\beta_t \in \{\hat{\beta}_t(\beta_{12}), \beta_{12} \in [\beta_{12}^l, \beta_{12}^u]\} = [\hat{\beta}_t(\beta_{12}^l), \hat{\beta}_t(\beta_{12}^u)] =: [\hat{\beta}_t^l, \hat{\beta}_t^u].
\]

The lower bound \( \hat{\beta}_t^l \) can be estimated from an OLS regression of \( Y - \beta_{12}^uT_{12} \) on \( X \). Similarly, the upper bound \( \hat{\beta}_t^u \) can be obtained from an OLS regression of \( Y - \beta_{12}^lT_{12} \) on \( X \). Under standard conditions, the OLS estimators \( \hat{\beta}_t^l \) and \( \hat{\beta}_t^u \) are asymptotically normal and the asymptotic variances \( A\text{var} (\hat{\beta}_t^l) \) and \( A\text{var} (\hat{\beta}_t^u) \) can be estimated consistently. We therefore apply the approach of Imbens & Manski (2004) and Stoye (2009) to construct confidence intervals for \( \beta_t \):\(^{33}\)

\[
CI_{1-\alpha} = \left[\hat{\beta}_t^l - c_{IM} \cdot \sqrt{\frac{A\text{var} (\hat{\beta}_t^l)}{N}}, \hat{\beta}_t^u + c_{IM} \cdot \sqrt{\frac{A\text{var} (\hat{\beta}_t^u)}{N}}\right],
\]

\(^{33}\)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).
where the critical value $c_{IM}$ solves
\[
\Phi \left( c_{IM} + \sqrt{N} \cdot \frac{\hat{\beta}^u_l - \hat{\beta}^l_l}{\sqrt{\max \left( \text{Avar} \left( \hat{\beta}^l_l \right), \text{Avar} \left( \hat{\beta}^u_l \right) \right)}} \right) - \Phi (-c_{IM}) = 1 - \alpha.
\]

Imbens & Manski (2004) and Stoye (2009) show that this is a valid confidence interval for $\beta_l$.

In Figure A.3, we report the rejection probabilities of a test that rejects if zero is not in the confidence interval (20). 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 $\beta_{12}$ can lead to power gains if the restrictions are correct but yield incorrect inferences if they are not.

(a) Size
(b) Power

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 $\alpha = 0.05$. 
### A.8 Additional figures and tables

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

<table>
<thead>
<tr>
<th></th>
<th>AER</th>
<th>ECMA</th>
<th>JPE</th>
<th>QJE</th>
<th>ReStud</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>1218</td>
<td>678</td>
<td>367</td>
<td>445</td>
<td>563</td>
<td>3271</td>
</tr>
<tr>
<td>Field experiment</td>
<td>43</td>
<td>9</td>
<td>14</td>
<td>45</td>
<td>13</td>
<td>124</td>
</tr>
<tr>
<td>Lab experiment</td>
<td>61</td>
<td>16</td>
<td>5</td>
<td>10</td>
<td>18</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>1322</td>
<td>703</td>
<td>386</td>
<td>500</td>
<td>594</td>
<td>3505</td>
</tr>
</tbody>
</table>
A.8.1 Ignoring the interaction

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

Bias

Size

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 curve

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.8.2 Pre-testing

**Figure A.3: Model selection: Bias, size, and power**

- **Bias**
  - Long
  - Short
  - Model selection

- **Size**
  - Long
  - Short
  - Model selection

- **Power**
  - Long
  - Short
  - Model selection

---

**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.8.3 Elliott et al. (2015)'s nearly optimal test

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

(a) Size

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 and Elliott et al. (2015)'s nearly optimal test: power curve

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.8.4 Armstrong et al. (2019)’s support restrictions on $\beta_{12}$

Figure A.6: Armstrong et al. (2019)’s support restrictions on $\beta_{12}$: size and power

(a) Size

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 and Armstrong et al. (2019)'s support restrictions on $\beta_{12}$: power curve

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 support restrictions on $\beta_{12}$.
A.8.5 Leaving the interaction cell empty

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

(a) Size

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.