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

Karthik Muralidharan† Mauricio Romero‡ Kaspar Wüthrich§

August 23, 2022

Abstract

Factorial designs are widely used to study multiple treatments in one experiment. While *t*-tests using a fully-saturated "long" model provide valid inferences, "short" model *t*-tests (that ignore interactions) yield higher power if interactions are zero, but incorrect inferences otherwise. Of 27 factorial experiments published in top-5 journals (2007–2017), 19 use the short model. After including interactions, over half of their results lose significance. Based on recent econometric advances, we show that power improvements over the long model are possible. We provide practical guidance for the design of new experiments and the analysis of completed experiments.

Keywords: randomized controlled trials; cross-cut designs; power in field experiments; data-dependent model selection; interaction effects; type-M errors

JEL Codes: C12, C18, C21, C90, C93

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

†Department of Economics, UC San Diego; NBER; J-PAL; E-mail: kamurali@ucsd.edu

‡Centro de Investigación Económica, ITAM, Mexico City, Mexico; J-PAL; E-mail: mtromero@itam.mx

§Department of Economics, UC San Diego; CESifo; Ifo Institute; E-mail: kwuthrich@ucsd.edu

1 Introduction

Cross-cutting or factorial designs are widely used in field experiments. For example, 27 out of 124 field experiments published in top-5 economics journals during 2007–2017 use cross-cutting designs. One rationale is that the power for detecting main treatment effects is higher if interactions between treatments are ignored in estimation and inference (with the implicit assumption that interactions are zero or negligible). This can make factorial designs a cost-effective way of studying multiple treatments.[1] A second rationale is to "explore" if there are meaningful interactions across treatments. This paper is motivated by the observation that both of these rationales can be problematic in practice.

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

The power gains from the short model, however, come at the cost of an increased likelihood of incorrect inference relative to a business-as-usual counterfactual (defined as outcomes in a pure experimental control group) if the interaction effect is not zero. Out of 27 field experiments published in top-5 economics journals during 2007–2017 using cross-cutting designs, 19 (over 70%) do not include all interaction terms in the main specifications. We reanalyzed the data from these papers by also including the interaction terms.[2] Doing so has non-trivial implications for inference on the main treatment effects. The median absolute value of the change in the point estimates is 96%, about 26% of estimates change sign, and 53% (29 out of 55) of estimates reported to be significant at the 5% level are no longer so after including interactions. Even if we reanalyze only "policy" experiments, 32% of the estimates (6 out of 19) are no longer significant after including interactions.[3]

In practice, researchers often estimate the long model first and test if the inter-

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

[2] The full list of 27 papers is in Table A.1. We reanalyzed 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.

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

action is significant, and then focus on the short model if they do not reject that the interaction is zero. However, such data-dependent model selection leads to invalid inferences (Leeb & Pötscher, 2005, 2006, 2008; Kahan, 2013) and should thus be avoided. Further, cross-cutting experiments are rarely adequately powered to detect meaningful interactions (see Section 2.6). Thus, this two-step procedure will almost always fail to reject that the interaction term is zero, even when it is different from zero. As a result, the rate of incorrect inference using this two-step model-selection procedure will continue to be nearly as high as that from just running the short model.

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

Textbook treatments of factorial designs (Cochran & Cox, 1957; Gerber & Green, 2012) and guides to practice (Kremer, 2003; Duflo et al., 2007) are careful to clarify that treatment effects using the short model should be interpreted as either (a) being conditional on the distribution of the other treatment arms in the experiment, or (b) as a composite treatment effect that includes a weighted-average of the interactions with other treatments. However, as we argue in Section 2.3, this weighted average is a somewhat arbitrary construct, can be difficult to interpret in high-dimensional factorial designs, and is typically neither of primary academic interest nor policy-relevant. Consistent with this view, none of the 19 experimental papers that focus on the short model motivate their experiment as being about estimating a weighted-average treatment effect.

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

Second, "absence of evidence" of significant interactions may be erroneously interpreted as "evidence of absence". The view that interactions are second-order (as

[4]For example, one would need an 8 times larger sample to detect an interaction than to detect a main effect when the interaction is half the size of the main effect; see Section 2.6 and Appendix A.3.

implied when papers only present the short model) may have been influenced partly by the lack of evidence of significant interactions in most experiments to date. However, as we show in Section 2.6, this is at least partly because few experiments are adequately powered to detect meaningful interactions. There is now both experimental (Duflo et al., 2015a; Mbiti et al., 2019) and non-experimental (Kerwin & Thornton, 2021; Gilligan et al., 2022) evidence that interactions matter. Indeed, a long tradition in development economics has highlighted the importance of complementarities across programs in alleviating poverty traps (Ray, 1998; Banerjee & Duflo, 2005), which suggests that assuming away interactions in empirical work may be a mistake.

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

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

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

Based on the analysis above, we recommend — in the interest of transparency

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

— that factorial experiments report results from the long regression model (even if only in an appendix). Long model t-tests are easy to compute even in complicated factorial designs and have appealing optimality properties. Further, the justification for omitting interactions should *not* be that these were not significant in the long model (because of the model selection issue discussed above). Rather, if researchers would like to focus on results from the short model, they should clearly indicate that treatment effects should be interpreted as composite effects that include a weighted-average of interactions with other treatments (and specify the estimand of interest in a pre-analysis plan). This will enable readers to assess the extent to which other treatments may be typical background factors that can be ignored.

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

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

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

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

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

each of these four rationales along with relevant examples in Section 5.

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

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

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

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

2 Factorial designs in theory

2.1 Setup

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

than two treatments.

Table 1: 2×2 factorial design

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

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

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

2.2 Long and short regression models

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

Long (or fully saturated) model:

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

Short model:

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

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

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

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

experimental control group) and the interaction effect:

$$\begin{aligned}
\beta_1 &= E\left(Y_{1,0} - Y_{0,0}\right) \quad \text{(ATE of } T_1 \text{ relative to a counterfactual where } T_2 = 0\text{)} && (3)\\
\beta_2 &= E\left(Y_{0,1} - Y_{0,0}\right) \quad \text{(ATE of } T_2 \text{ relative to a counterfactual where } T_1 = 0\text{)} && (4)\\
\beta_{12} &= E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right) \quad \text{(interaction effect)}^{7} && (5)
\end{aligned}$$

By contrast, the regression coefficients in the short model are

$$\begin{aligned}
\beta_1^s &= E\left(Y_{1,1} - Y_{0,1}\right) P\left(T_2 = 1\right) + E\left(Y_{1,0} - Y_{0,0}\right) P\left(T_2 = 0\right) && (6)\\
&= E\left(Y_{1,0} - Y_{0,0}\right) + E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right) P(T_2 = 1) && (7)\\
\beta_2^s &= E\left(Y_{1,1} - Y_{1,0}\right) P\left(T_1 = 1\right) + E\left(Y_{0,1} - Y_{0,0}\right) P\left(T_1 = 0\right) && (8)\\
&= E\left(Y_{0,1} - Y_{0,0}\right) + E\left(Y_{1,1} - Y_{0,1} - Y_{1,0} + Y_{0,0}\right) P(T_1 = 1) && (9)
\end{aligned}$$

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

Remark 1. *The problem of choosing between the long model and the short model is not unique to factorial designs and arises in many contexts. For example, when estimating treatment effects in observational studies, researchers need to decide whether to include the covariates linearly or consider fully interacted specifications (e.g., Angrist & Krueger, 1999; Angrist & Pischke, 2009). However, the practical implications are not the same because experimental treatments are fundamentally different in nature from standard covariates, as we discuss below in Section 2.3. The choice between the short and the long model (with interactions between the treatment and strata indicators) is also relevant in stratified experiments (e.g., Imbens & Rubin, 2015; Ansel et al., 2018; Bugni et al., 2018, 2019).*

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

Section 2.2 shows that the short model yields a weighted average of treatment effects that depends on the nature and distribution of the other treatment arms in the experiment. This weighted average is typically neither of primary academic interest

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

nor policy-relevant. This view is consistent with how papers we reanalyze motivate their object of interest, which is usually the main treatment effect against a business-as-usual counterfactual. Of the 19 papers in Table A.1 in Appendix A.1 that present results from the short model without all interactions, we did not find any study that mentioned (in the main text or a footnote) that the presented treatment effects should be interpreted as either (a) a composite effect that includes a weighted average of the interaction with the other treatments or (b) as being against a counterfactual that was not business-as-usual but one that also had the other treatments in the same experiment.

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

However, experimental treatments are fundamentally different from standard background covariates. They are determined by the experimenter based on research interest, and rarely represent real-world counterfactuals. In some cases, the interventions studied are new and the other treatments may not even exist in the study population. Even if they do exist, there is no reason to believe that the distributions in the population mirror those in the experiment. Thus, we view this issue as a challenge to internal validity. Further, papers with factorial designs often use the two-step procedure described in Section 2.5, and present results from the short model *after* mentioning that the interactions are not significantly different from zero (e.g., Banerjee et al., 2007; Karlan & List, 2007). This suggests that our view that interactions matter for internal validity is shared broadly.

Finally, even in settings where the coefficients in the short model are of interest, they can always be constructed based on the coefficients in the long model, while the converse is not true. One can also use the long model to test hypotheses about the coefficients in the short regression model: $H_0 : \beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1) = 0$. Which test is more powerful depends on the relative sample size in the four experimental cells.[8] Unlike the short model, the long model additionally allows for testing a rich variety of hypotheses about counterfactual effects such as $H_0 : \beta_1 + \beta_{12} p = 0$ for policy-relevant values of p, which generally differ from the experimental assignment probability $P(T_2 = 1)$. For instance, resume audit experiments may vary charac-

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

teristics such as age, gender, race, education, and experience with the sample size allocated to various combinations of these characteristics being different from their proportion in the population. In such a case, short model estimates are difficult to interpret, whereas estimating the long model and calculating a weighted average of main and interaction effects with weights equal to their population proportions may yield a more policy-relevant treatment effect.

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

2.4 Inference on main effects

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

To illustrate, suppose the data generating process is given by

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

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

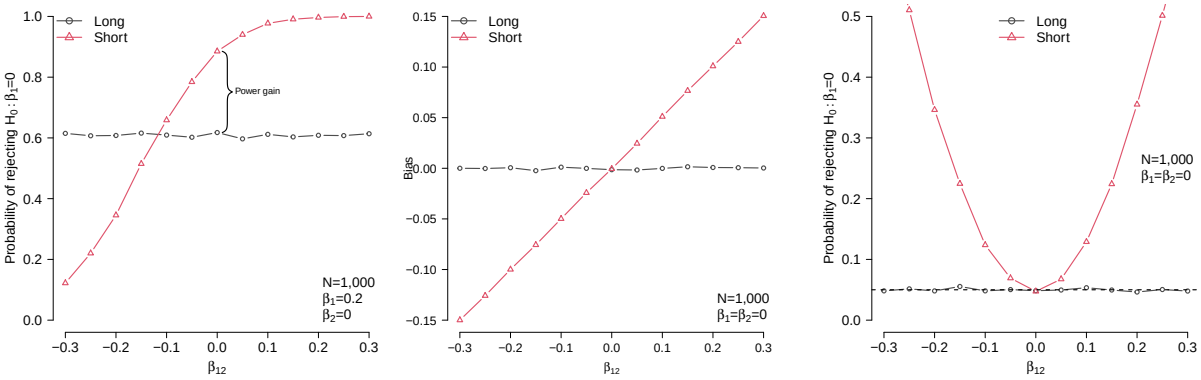
If, on the other hand, $\beta_{12} \neq 0$, ignoring the interaction can lead to substantial size distortions. To illustrate, we introduce a simple running example. Consider a 2×2 design with a total sample size of $N = 1,000$ and $N_1 = N_2 = N_3 = N_4 = 250$. The data are generated based on Model (10)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

Figure 1 shows how power, bias, and size vary across different values of β_{12} in both the long and the short model. When $\beta_{12} = 0$, the short model t-test controls size and exhibits higher power than the long model t-test as discussed before. However, these power gains come at the cost of bias and size distortions whenever $\beta_{12} \neq 0$. Importantly, even modest values of $|\beta_{12}|$ lead to considerable size distortions. For instance, $|\beta_{12}| > 0.1\sigma$ more than doubles the rate of false rejection of the null (in the data we reanalyze in Section 3.2, we find that $|\hat{\beta}_{12}| > 0.1\sigma$ in over 36% of cases). By contrast, the long model is unbiased and exhibits correct size for all values β_{12}. The

main takeaway from Figure 1 is that researchers should avoid the short model for making inference on the main effects, unless they are certain that $\beta_{12} = 0$.

Figure 1: The perceived power gains from the short model come at the cost of biased estimators and not controlling size, unless β_{12} is exactly equal to zero

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



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

2.5 Model selection (or pre-testing) yields invalid inferences

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

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

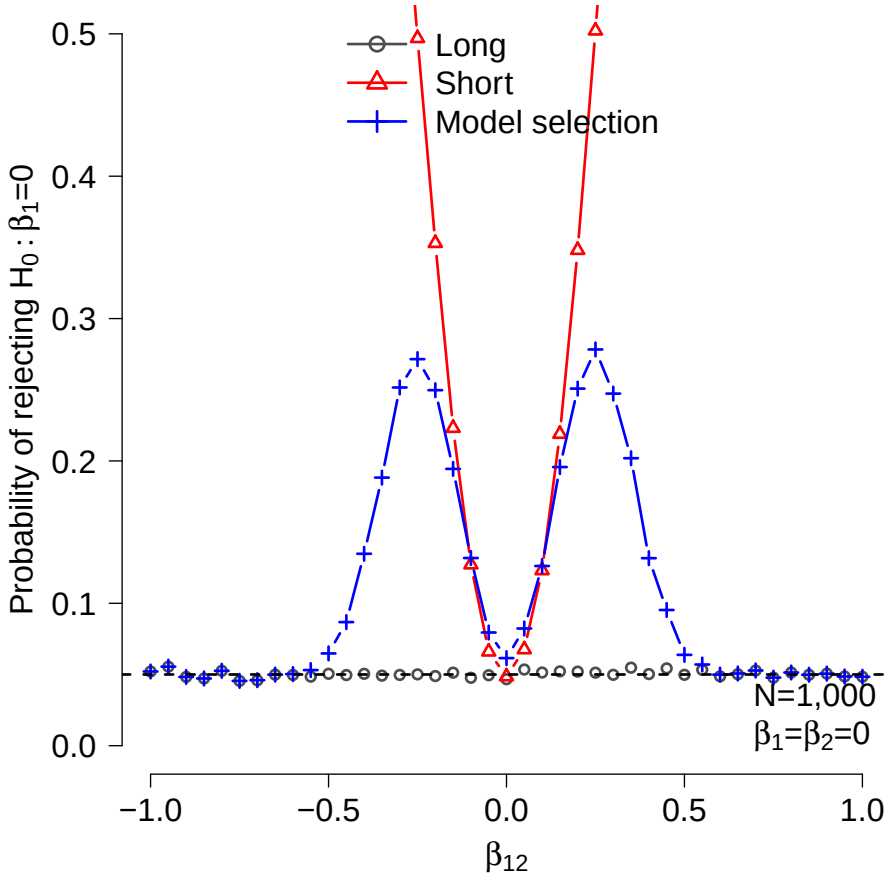
While seemingly attractive, such data-dependent model selection leads to invalid inferences (e.g., Leeb & Pötscher, 2005, 2006, 2008; Kahan, 2013). Figure 2 shows the size properties of the two-step model selection approach in our running example. For reference, we also include results for the short and long model t-tests. The main takeaway from Figure 2 is that model selection leads to incorrect inferences and false positives for a wide range of values of β_{12}.[9] Model selection can be particularly

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

problematic for program evaluation field experiments because they are expensive to run, and therefore typically not adequately powered to reject that the interactions are zero (Section 2.6).

The range of values for $|\beta_{12}|$ for which model selection leads to substantial size distortions shrinks as the sample size (and power) of the experiment increases. However, it can be quite large in realistic settings. In our running example, with 1,000 observations one would need $|\beta_{12}|$ to be above 0.5 to avoid notable size distortions. Even with 10,000 observations, only values of $|\beta_{12}|$ above 0.2 lead to negligible size distortions (see Figure A.13). Since the true value of the interaction is unknown and likely to be in this "problematic range" in many practical settings (see Figure 3), we recommend that researchers avoid the data-driven model-selection approach.

Figure 2: Model selection does not control size



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

Remark 2. *As Figure 2 shows, model selection is less of a concern when the interactions are either zero or very large, but is a first-order issue when interactions are in the problematic*

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

2.6 Inference on interaction effects

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

However, detecting interaction effects requires much larger sample sizes than needed for detecting main effects. To illustrate, we compare the standard errors of the OLS estimator of the interaction effect, $\hat{\beta}_{12}$, and the main effect, $\hat{\beta}_1$. Under the assumptions in Section 2.4, the standard errors are $SE\left(\hat{\beta}_1\right) = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$ and $SE\left(\hat{\beta}_{12}\right) = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4}}$. Since $SE\left(\hat{\beta}_1\right) < SE\left(\hat{\beta}_{12}\right)$, the power for detecting interaction effects is always lower than the power for detecting main effects, and the required sample size for detecting interaction effects is always larger than the required sample size for detecting main effects of equal magnitude. For example, we need eight times the sample size to have the same power to detect an interaction effect as to detect the main effect, when the interaction is half the size of the main effect (see Appendix A.3). Given the more onerous sample size requirements to detect interactions relative to main effects, it is not surprising that only few of the interaction effects are significant in the reanalysis in Section 3.2.1.

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

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

significant. Figure A.9 in Appendix A.3 shows the relationship between the Type-M error and the power of the experiment.

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

3 Factorial designs in practice

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

3.1 Data and descriptive statistics

We analyze all articles published between 2007 and 2017 in the top five journals in Economics.[11] Of the 3,505 articles published in this period, 124 (3.5%) are field experiments (Table A.6 provides more details). Factorial designs are widely used: Among 124 field experiments 27 (22%) had a factorial design.[12] Only 8 of these 27 articles with factorial designs (∼30%) used the long model including all interaction terms as their main specification (see Table 2).

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

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

Table 2: Field experiments published in top-5 journals between 2007 and 2017

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

3.2 Ignoring interactions in practice

In Section 2.4, we have shown that ignoring interactions can lead to substantial size distortions and false positives. Here, we examine the practical implications of ignoring the interactions in the papers listed in Table A.1. We reanalyze the data from all field experiments with factorial designs and publicly available data that do not include all the interactions in the main specification.[13] Of the ten most-cited papers with factorial designs listed in Table 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 (over 70%).[14] Of these 19, 4 papers did not have publicly-available replication data. In an online appendix we describe the experimental design of each of the 27 papers and provide details on our replication analysis.[15]

We downloaded the publicly-available data files and replicated the main results in each of the remaining 15 papers. We standardized the outcome variable in each paper to have mean zero and standard deviation of one. We then compared the original treatment effects (estimated without the interaction terms) with those estimated including the interaction terms. In other words, we compare estimates based on the short model (Equation (2)) to those based on the long model (Equation (1)).

3.2.1 Key facts about interactions

As the discussion in Section 2.4 highlights, the extent to which the short model will not control size depends on the value of the interactions in practice. We therefore

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

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

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

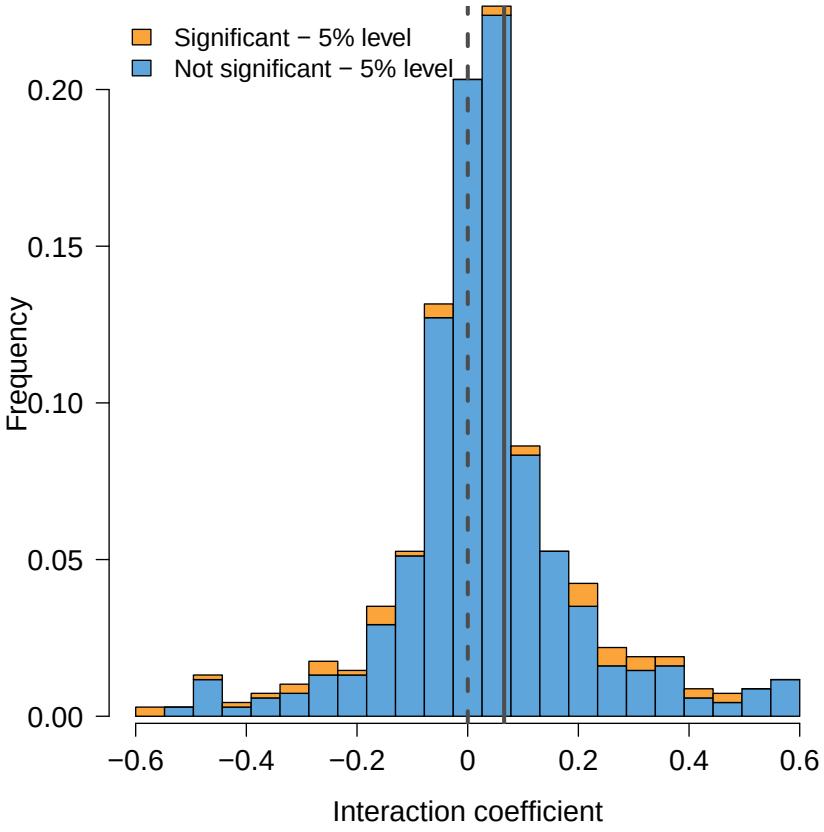
start by plotting the distribution of estimated interaction effects (Figure 3) and documenting facts regarding interactions from our reanalysis. We find that interactions are quantitatively important and typically not second-order. All estimates are measured in standard deviations (σ) of the outcome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

The second key finding is that most experiments will rarely reject the null hypothesis that the interactions are zero (Figure 3 shades the fraction of the interactions that are significant in the studies that we reanalyze). Among the 15 papers that we reanalyzed, 6.2% of interactions (spread across 4 papers) are significant at the 10% level, 3.6% are significant at the 5% level (spread across 3 papers), and 0.9% are significant at the 1% level (in 1 paper).[16] These findings are not surprising because factorial designs are rarely powered to detect meaningful interactions.

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

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

Figure 3: Distribution of the estimated interaction effects



Note: This figure shows the distribution of the interactions between the main treatments (N=868 in this figure). We trim the top and bottom 1% of the distribution. The median interaction for these papers is 0.00σ (dashed vertical line), the median absolute value of the interaction is 0.07σ (solid vertical line), and the median relative absolute value of the interaction with respect to the main treatment effect is 0.37. 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 Ignoring interactions has important implications for estimation and inference

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

Table 3 shows how the significance of the main treatment estimates changes when using the long instead of the short model. About 48% of treatment estimates that were significant at the 10% level based on the short model are no longer significant

based on the long model. 53% and 57% of estimates lose significance at the 5% and 1% levels, respectively. A much smaller fraction of treatment effects that were not significant in the short model are significant based on the long regression (6%, 5%, and 1%, at the 10%, 5%, and 1% levels, respectively).[17]

We find similar results when we restrict our reanalysis to the ten most cited papers with factorial designs that do not include the interaction terms (with data available for reanalysis). When we re-estimate the treatment effects in these papers after including interactions, we find that out of 21 results that were significant at the 5% level in the paper, 9 (or 43%) are no longer so after including interactions. Corresponding figures and tables are presented in Appendix A.1.2 (Figure A.2 and Table A.2).

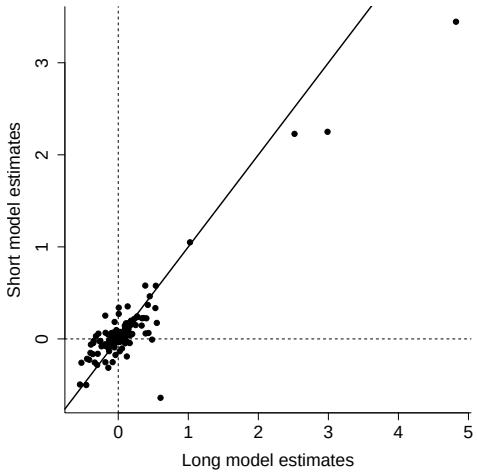
Finally, we also distinguish between policy and conceptual experiments in Table A.1 (the latter typically have more treatments and interactions) and see that the problem of incorrect inference from ignoring interaction terms remains even when we restrict attention to the policy experiments. Of the 12 policy experiments, 9 do not include all interactions. When we re-estimate the treatment effects in these 9 papers after including 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 interactions. Corresponding figures and tables are presented in Appendix A.1.3 (Figure A.4 and Table A.3).[18]

[17]These results are not driven by just a few papers. If we first estimate the median change in the absolute value of the estimate *within* each paper, and then the median change across papers, the result is similar to estimating the median absolute changes across all estimates at 97%. Likewise, if we first estimate the proportion of estimates that change sign within each paper, and then estimate the average across papers the result is 25%, which is similar to estimating the proportion of estimates that change sign. Finally, 73% of papers have at least one estimate that is no longer significant at the 10% level when estimating the full model, 77% have at least one estimate that is no longer significant at the 5% level, and 82% have at least one estimate that is no longer significant at the 1% level.

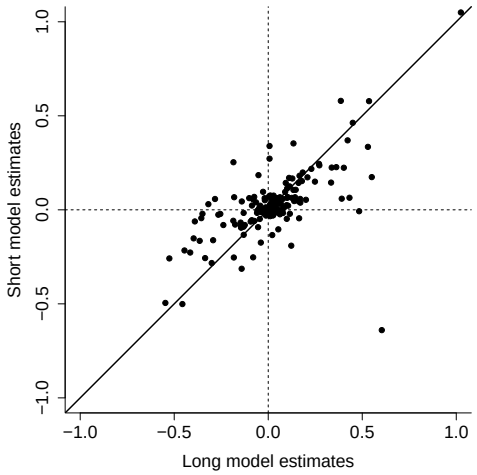
[18]Among the papers that originally included all interactions, 23% of results that are significant at the 5% level in the short model are not significant in the long model. See Appendix A.1.4 for more details.

Figure 4: Treatment effects estimates based on the long and the short model

(a) All estimates



(b) Main treatment effects between -1σ and 1σ



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

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

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

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

4 Improving power for detecting main effects

We now examine whether it is possible to improve power for detecting main effects relative to long model t-tests, while maintaining size control for relevant values of the interactions. We consider 2×2 factorial designs and briefly comment on factorial designs with more than two treatments at the end of each subsection. Throughout, we will focus on the main ideas underlying the different econometric methods. Appendix A.4 provides detailed descriptions and implementation details.

4.1 Setup

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

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

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

The two-sided long model t-test is the uniformly most powerful test among tests that are unbiased for all values of the interaction effect (e.g., van der Vaart, 1998; Elliott et al., 2015).[19] This implies that any test that is more powerful than the long model t-test for some values of β_{12} must have lower power somewhere else. Thus, to achieve higher power than the long model t-test, one has to choose which values of β_{12} to direct power to based on prior knowledge.

If one insists on size control for all β_{12}, the scope for power improvements relative to the long model t-test is theoretically limited.[20] For example, at the 5%-level, the maximal theoretically possible power improvement over the long model two-sided t-test is 12.5 percentage points. Section 4.2 proposes a nearly optimal test that comes close to achieving the maximal power gain at a priori likely values of the interaction, while controlling size for all values of the interaction. In Appendix A.6, we show that a Bonferroni-style correction after model selection leads to local power improvements for a range of positive values of the interaction.

The limited scope for power improvements relative to the long model t-test motivates relaxing the uniform size control requirement and imposing additional restrictions on β_{12}. An extreme example is the short model t-test, which can improve power relative to long model t-test by much more than 12.5%, but only controls size under the restrictive assumption that $\beta_{12} = 0$. In Section 4.3, we explore an intermediate approach that restricts the magnitude of β_{12}, which is often more realistic than assuming that β_{12} is exactly equal to zero.

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

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

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

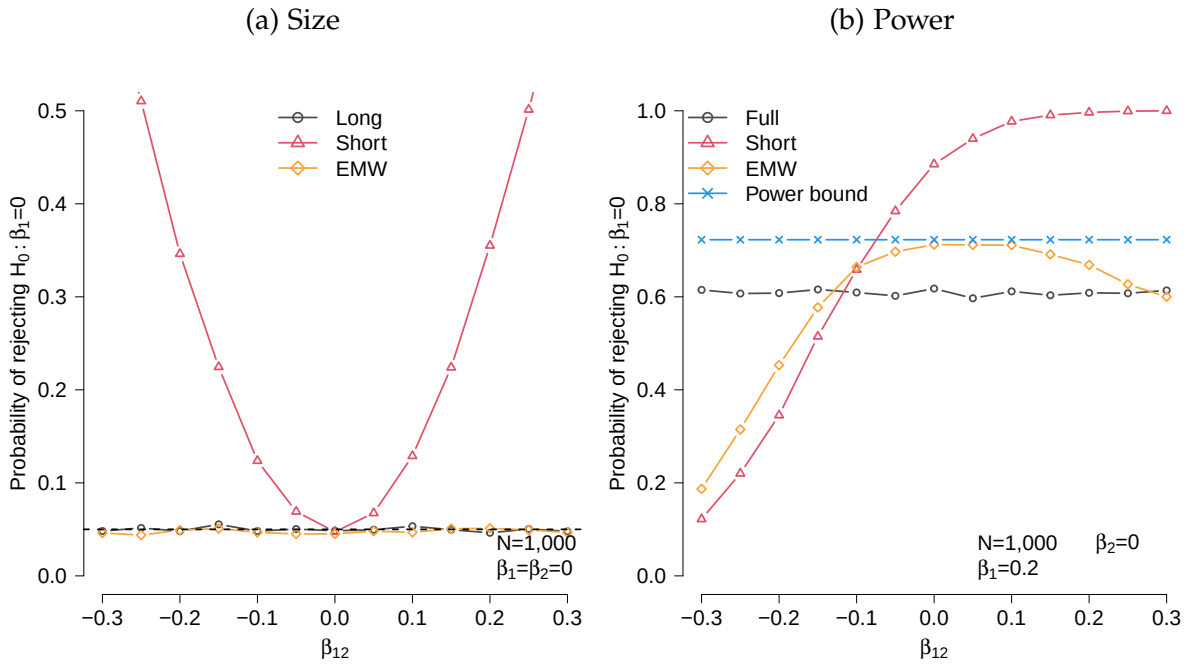
[20] This is because the one-sided long model t-tests are uniformly most powerful (e.g., Proposition 15.2 in van der Vaart, 1998) so that, for any β_{12}, the maximal power is achieved by a one-sided t-test (e.g., Armstrong & Kolesar, 2015, 2021). See Armstrong & Kolesar (2018) for a discussion of the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

testing problem

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

We use the numerical algorithm developed by Elliott et al. (2015) to construct a nearly optimal test for the testing problem (12).[21] Elliott et al. (2015) consider a setting where one is interested in maximizing weighted average power. The best test in this setting is a Neyman-Pearson test based on the least favorable distribution (LFD). Since the LFD is often difficult to compute analytically, Elliott et al. (2015) instead focus on an approximate LFD, which yields feasible and nearly optimal tests.

Figure 5: Elliott et al. (2015)'s nearly optimal test controls size and yields power gains over running the full model near $\bar{\beta}_{12} = 0$



Note: Simulations are based on the running example with sample size N, normal iid errors, and 10,000 repetitions. The size for Figures 5a and 5b is $\alpha = 0.05$. EMW refers to Elliott et al. (2015)'s nearly optimal test. The power bound in Figure 5b is the power of the one-sided long model t-test for the testing problem $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$.

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

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

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

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

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

If the researcher is certain that $\beta_{12} = \bar{\beta}_{12}$, 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 the short model t-test. As shown in Section 2.4, short model t-tests are more powerful than long model t-tests when $\beta_{12} = 0$, but do not control size when $\beta_{12} \neq 0$.

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

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

Assumption 1 restricts the parameter space for β_{12} and implies that $\beta_{12} \in [-C, C]$. We explore two different approaches for making inferences under this assumption. First, we construct optimal confidence intervals under Assumption 1 based on the approach developed by Armstrong et al. (2020). Their confidence intervals are based on linear estimators for β_1 and account for the worst case bias of the estimators. As a result, the length of the confidence interval is determined by the bias and the variance of the estimator, and to obtain optimal confidence intervals one has to solve a bias-variance trade-off. This problem can be solved using convex optimization. We refer to this approach as the Armstrong-Kolesar-Kwon (AKK) approach.

The second approach is based on constructing bounds on the main effect implied by Assumption 1. In particular, upper and lower bounds on β_1 can be obtained from regressions of $Y + CT_{12}$ on T_1 and $Y - CT_{12}$ on T_1, respectively. We apply the procedure of Imbens & Manski (2004) and Stoye (2009) to construct valid confidence intervals for β_1. We refer to this approach as the Imbens-Manski-Stoye (IMS) approach.[22]

[22] As outlined in Appendix A.4.3, it is straightforward to use the IMS approach if the prior information takes the form $C_1 \leq \beta_{12} \leq C_2$ for any $-\infty < C_1 < C_2 < \infty$, which may be more appropriate in some settings. Further, one could make inferences under restrictions on the direction of the interaction effects using the approach by Ketz & McCloskey (2021). Both types of approaches may be suitable in cases where there is a strong prior that treatments are complements or substitutes.

In Figure 6, we report the rejection probabilities of tests that reject if zero is not in the AKK and IMS confidence intervals. To illustrate, we assume that $C = 0.1$, implying that $\beta_{12} \in [-0.1, 0.1]$.[23] Our results suggest that AKK and IMS can be substantially more powerful than long model t-tests when the prior knowledge is correct, but may exhibit size distortions when it is not. Panel (b) shows that the AKK and IMS power curves cross at zero. Thus, the choice between the two approaches should be based on which values of the interaction the researchers want to direct power to. Appendices A.7.4 and A.7.5 present the corresponding power curves for different values of β_1.

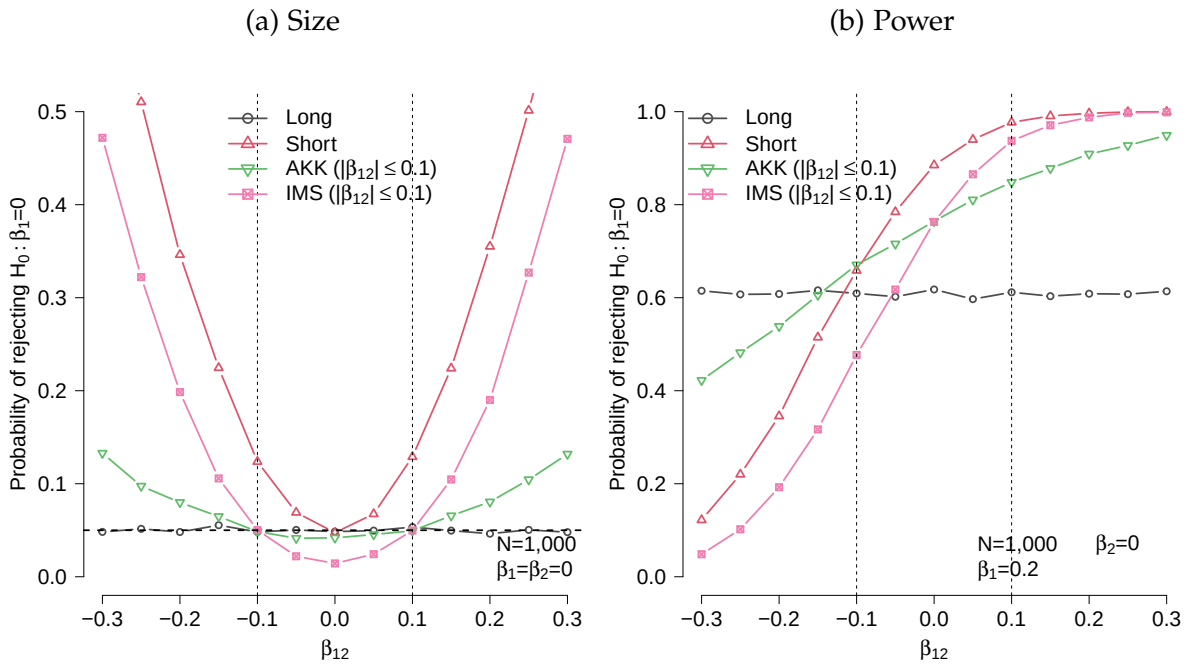
When researchers are primarily interested in the main effects and feel confident that the interactions are second-order, AKK and IMS should be strictly preferred to the short model, since it is more realistic to pre-specify that the interaction is in a range than exactly zero. However, pre-specifying the appropriate range of prior values for the interaction is non-trivial and requires judgment.[24]

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

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

[24] It is problematic to use AKK or IMS based on first running the long model and not rejecting that the interaction is in a certain range. This would result in data-dependent model selection issue similar to those documented in Section 2.5. Thus, while AKK and IMS are improvements over the short model, they do not solve the underlying problem of not knowing the true value of the interaction.

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



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

4.4 A design-based approach for improving power

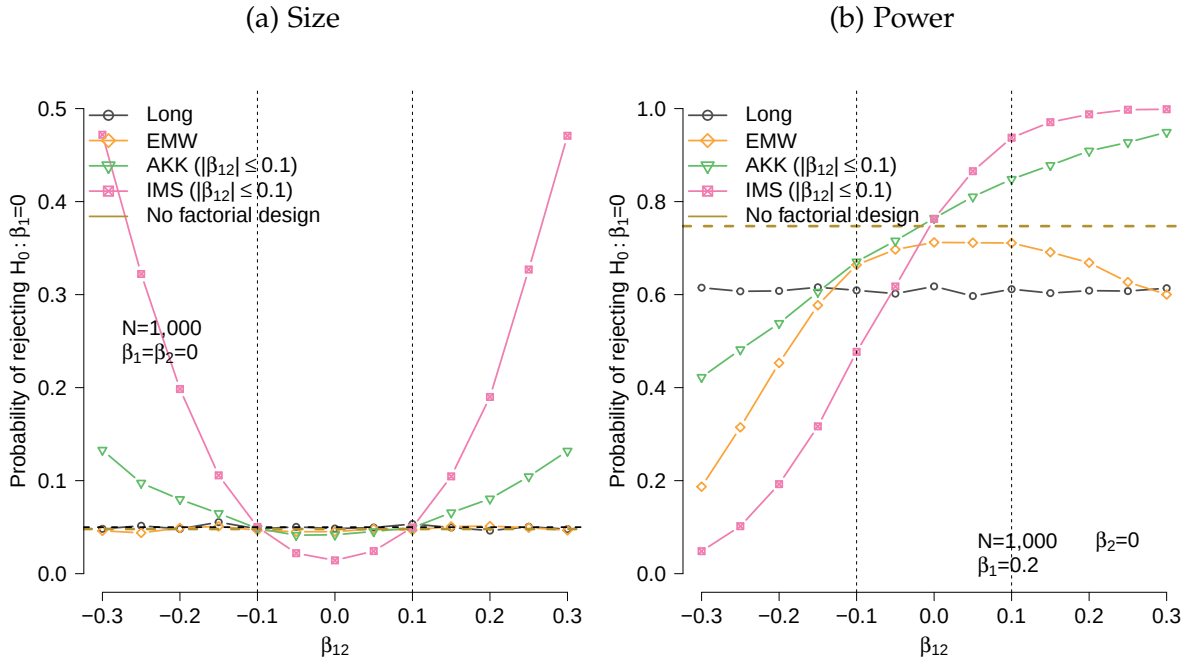
The discussion above focused on improving power for detecting main effects in existing experiments with factorial designs. While these techniques can also be used to analyze new experiments (and be included in a pre-analysis plan), a design-based alternative is to leave the "interaction cell" empty (i.e., to set $N_4 = 0$) and to re-assign those subjects to the other cells (see Table A.5).

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

This design (with interaction cells empty) yields power gains relative to running two separate experiments, because the control group is used twice. But it avoids the problem of interactions discussed above. An example of such a design is provided

by Muralidharan & Sundararaman (2011) who study the impact of four different interventions in one experiment with one common control group, but no cross-cutting treatment arms.

Figure 7: Leaving the interaction cell empty increases power relative to approaches that control size for all β_{12}



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

4.5 Which econometric approach should one use in practice?

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

For the reanalysis of existing experiments, the choice of the econometric method for making inferences on the main effects should be based on the strength of the available prior knowledge. If researchers have little prior knowledge about the interaction

effects, we recommend using the long model t-tests, which are the uniformly most powerful unbiased tests. If prior knowledge about the interaction effects is available, but the researchers are not confident enough to be willing to sacrifice size control for all values of the interactions, we recommend Elliott et al. (2015)'s nearly optimal tests. The nearly optimal test allows for targeting power based on prior knowledge while ensuring uniform size control. If precise prior knowledge about the interaction effects is available, researchers can use the AKK or the IMS approach to leverage such prior knowledge to improve power substantially. However, unlike the other methods, these two approaches exhibit size distortions when the prior knowledge is incorrect.

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

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

5 When does the short model make sense?

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

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

factorial designs an efficient way to quickly learn about which of several treatment dimensions that could be manipulated may be worth studying and testing further. In contrast, policy experiments are typically run only once, making factorial designs and short model estimates less desirable.

The second case is when the goal of the experiment is not hypothesis testing but to minimize MSE criteria (or other loss functions), which involve a bias-variance trade-off in estimating the main effects. For example, for small values of the interaction effects, estimators based on the short model can yield a lower root MSE than the estimators based on the design which leaves the interaction cell empty (Blair et al., 2019). These alternative criteria also justify the use of factorial designs for agricultural experiments and online A/B testing, since their goal is to optimize decision-making over several factors (to maximize yields or profits) as opposed to testing if individual factors are "significant". Again, this contrasts with the case of policy experiments, where the goal is typically to test if a program or policy had a significant effect, and factorial designs and short-model inferences may therefore be problematic.

The third case is to improve an experiment's external validity. Cochran & Cox (1957, p.152) recommend factorial designs for "experiments designed to lead to recommendations that must apply over a wide range of conditions. Subsidiary factors may be brought into an experiment so as to test the principal factors under a variety of conditions similar to those that will be encountered in the population to which recommendations are to apply"; see also the discussion in Fisher (1992). Thus, factorial designs and the short model may be fine when one dimension of the experiment is studying reasonable variants of the main treatment, but less so when all treatments are of primary interest.[25]

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

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

6 Conclusion

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

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

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

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

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

Finally, it is worth noting that factorial designs *do* provide an efficient way of learning about multiple treatments as well as their interactions in the same experiment.

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

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights*, *2*(2), 193-208.

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

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

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

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

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

Angrist, J. D., & Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), (Vol. 3, p. 1277 - 1366). Elsevier.

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

Ansel, J., Hong, H., & Li, J. (2018). OLS and 2SLS in randomized and conditionally randomized experiments. *Jahrbücher für Nationalökonomie und Statistik*, *238*(3-4), 243–293.

Armstrong, T. B., & Kolesar, M. (2015). *Optimal inference in a class of regression models.* arXiv:1511.06028v2. Retrieved from https://arxiv.org/abs/1511.06028

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

Armstrong, T. B., & Kolesar, M. (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, *12*(1), 77-108.

Armstrong, T. B., Kolesar, M., & Kwon, S. (2020). *Bias-aware inference in regularized regression models.* arXiv:2012.14823.

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

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73–140). Elsevier.

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

Banerjee, A., Chandrasekhar, A. G., Dalpath, S., Duflo, E., Floretta, J., Jackson, M. O., ... Shrestha, M. (2021). *Selecting the most effective nudge: Evidence from a large-scale experiment on immunization* (Working Paper No. 28726). National Bureau of Economic Research.

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

122(3), 1235-1264.

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

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

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

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

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634-60.

Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32.

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

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

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *113*(524), 1784-1796.

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, *10*(4), 1747-1785.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920-80.

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

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

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

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

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

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

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

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

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

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

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

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

Fisher, R. A. (1992). The arrangement of field experiments. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 82–91). New York, NY: Springer New York.

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

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

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

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

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

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

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A. M., & Neal, D. (2022). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources*, *57*(1), 79–111.

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

Hansen, B. E. (2022). *Econometrics*. Princeton University Press.

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

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*(6), 1845–1857.

Imbens, G. W., & Rubin, D. B. (2015). Stratified randomized experiments. In *Causal inference for statistics, social, and biomedical sciences: An introduction* (pp. 187–218). Cambridge University Press.

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

Kahan, B. C. (2013). Bias in randomised factorial trials. *Statistics in medicine*, *32*(26), 4540–4549.
Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774-1793.
Karlan, D., Osei, R., Osei-Akoto, I., & Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, *129*(2), 597–652.
Karlan, D., & Zinman, J. (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, *98*(3), 1040-68.
Karlan, D., & Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, *77*(6), 1993–2008.
Kassler, D., Nichols-Barrer, I., & Finucane, M. (2019). Beyond treatment versus control: How bayesian analysis makes factorial experiments feasible in education research. *Evaluation Review*.
Kaur, S., Kremer, M., & Mullainathan, S. (2015). Self-control at work. *Journal of Political Economy*, *123*(6), 1227–1277.
Kendall, C., Nannicini, T., & Trebbi, F. (2015). How do voters respond to information? evidence from a randomized campaign. *American Economic Review*, *105*(1), 322-53.
Kerwin, J. T., & Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, *103*(2), 251-264.
Ketz, P., & McCloskey, A. (2021). *Short and simple confidence intervals when the directions of some effects are known.* Working paper.
Khan, A. Q., Khwaja, A. I., & Olken, B. A. (2015). Tax farming redux: Experimental evidence on performance pay for tax collectors. *The Quarterly Journal of Economics*, *131*(1), 219–271.
Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in Denmark. *Econometrica*, *79*(3), 651-692.
Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.
Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *The American Economic Review*, *93*(2), pp. 102-106.
Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21-59.
Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 2554–2591.
Leeb, H., & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(02), 338–376.
Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.
List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439.
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1–21.
Lu, J., Qiu, Y., & Deng, A. (2019). A note on Type S/M errors in hypothesis testing.

British Journal of Mathematical and Statistical Psychology, *72*(1), 1-17.
Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *The Quarterly Journal of Economics*, *134*(3), 1627-1673.
McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*.
McCloskey, A. (2020). Asymptotically uniform tests after consistent model selection in the linear regression model. *Journal of Business & Economic Statistics*, *38*(4), 810-825.
Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.
Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, *115*(2), 200-249.
Pallais, A., & Sands, E. G. (2016). Why the referential treatment? evidence from field experiments on referrals. *Journal of Political Economy*, *124*(6), 1793–1828.
Ray, D. (1998). *Development economics*. Princeton University Press.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.
Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, *77*(4), 1299–1315.
Thornton, R. L. (2008). The demand for, and impact of, learning HIV status. *American Economic Review*, *98*(5), 1829-63.
van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.
Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133.
Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond $p < 0.05$. *The American Statistician*, *73*(sup1), 1-19.
Young, A. (2018). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics*, *134*(2), 557-598.

A Online appendix for "Factorial designs, model selection, and (incorrect) inference in randomized experiments"

A.1 Papers with factorial designs published in Top-5 economics journals

Table A.1: Papers with factorial designs published between 2007 and 2017 in top-5 economics journals sorted by citation count (as of July 4, 2019)

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Olken (2007)	Monitoring Corruption: Evidence from a Field Experiment in Indonesia	JPE	2007	1529	3	2	0	Yes	Yes
Banerjee et al. (2007)	Remedying Education: Evidence from Two Randomized Experiments in India	QJE	2007	1213	2	1	0	Yes	Yes
Duflo et al. (2011)	Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya	AER	2011	787	3	4	0	Yes	Yes
Kleven et al. (2011)	Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark	ECMA	2011	776	2	1	0	No	Yes
Karlan et al. (2014)	Agricultural Decisions after Relaxing Credit and Risk Constraints	QJE	2014	612	2	1	1	No	Yes

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Bertrand et al. (2010)	What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment	QJE	2010	522	14	85	0	Yes	No
Karlan & List (2007)	Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment	AER	2007	506	7	28	0	Yes	No
Thornton (2008)	The Demand for, and Impact of, Learning HIV Status	AER	2008	453	2	1	0	Yes	Yes
Haushofer & Shapiro (2016)	The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya	QJE	2016	393	6	8	3	Yes	Yes
Alatas et al. (2012)	Targeting the Poor: Evidence from a Field Experiment in Indonesia	AER	2012	330	4	16	0	Yes	Yes
Karlan & Zinman (2008)	Credit Elasticities in Less-Developed Economies: Implications for Microfinance	AER	2008	311	3	2	0	Yes	No
Duflo et al. (2015a)	Education, HIV, and Early Fertility: Experimental Evidence from Kenya	AER	2015	282	3	3	1	Yes	Yes
Andreoni et al. (2017)	Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving	JPE	2017	270	2	1	1	Yes	No

Continued on next page

Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Jakiela & Ozier (2015)	Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies	ReStud	2016	245	3	6	6	Yes	No
Eriksson & Rooth (2014)	Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment	AER	2014	238	34	71680	0	Yes	No
Allcott & Taubinsky (2015)	Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market	AER	2015	237	2	1	0	No	No
Flory et al. (2014)	Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions	ReStud	2015	204	10	24	12	Yes	No
Brown et al. (2010)	Shrouded Attributes and Information Suppression: Evidence from the Field	QJE	2010	189	3	6	6	No	No
DellaVigna et al. (2016)	Voting to Tell Others	ReStud	2017	169	4	15	0	Yes	No
Fischer (2013)	Contract Structure, Risk-Sharing, and Investment Choice	ECMA	2013	162	7	9	9	Yes	No
Kaur et al. (2015)	Self-Control at Work	JPE	2015	154	8	16	0	Yes	No

Continued on next page

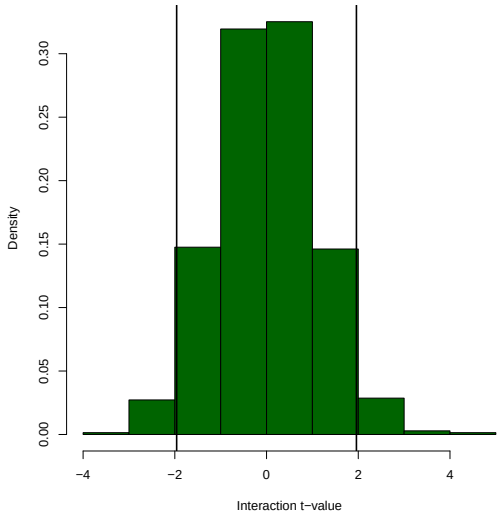
Table A.1 – continued from previous page

Authors	Title	Journal	Year	Citations	Treatments	Interactions In Design	Interactions Included	Data Available	Policy Evaluation
Cohen et al. (2015)	Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial	AER	2015	151	3	7	7	Yes	Yes
Blattman et al. (2017)	Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia	AER	2017	135	2	1	1	Yes	Yes
Khan et al. (2015)	Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors	QJE	2016	133	6	8	0	Yes	Yes
Balafoutas et al. (2013)	What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods	ReStud	2013	126	5	6	0	Yes	No
Kendall et al. (2015)	How Do Voters Respond to Information? Evidence from a Randomized Campaign	AER	2015	116	5	5	5	Yes	No
Pallais & Sands (2016)	Why the Referential Treatment? Evidence from Field Experiments on Referrals	JPE	2016	85	3	12	0	No	No

Note: This table provides relevant information from all articles with factorial designs published in top-5 journals. Citation counts are from Google Scholar on July 4th of 2019. Treatments is the number of different treatments in the paper. "Interactions in Design" is the number of interactions in the experimental design. "Interactions Included" is the number of interactions included in the main specification of the paper. Data available, refers to whether the data is publicly available or not. Allcott & Taubinsky (2015) has two field experiments. The table refers to the second one. One of the three dimensions of randomization in Flory et al. (2014) does not appear in the publicly available data. Online Appendix B (in http://mauricio-romero.com/pdfs/papers/Appendix_crosscuts.pdf) describes the experimental design of each of the 27 papers and provides further details on our replication analysis.

A.1.1 All papers

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

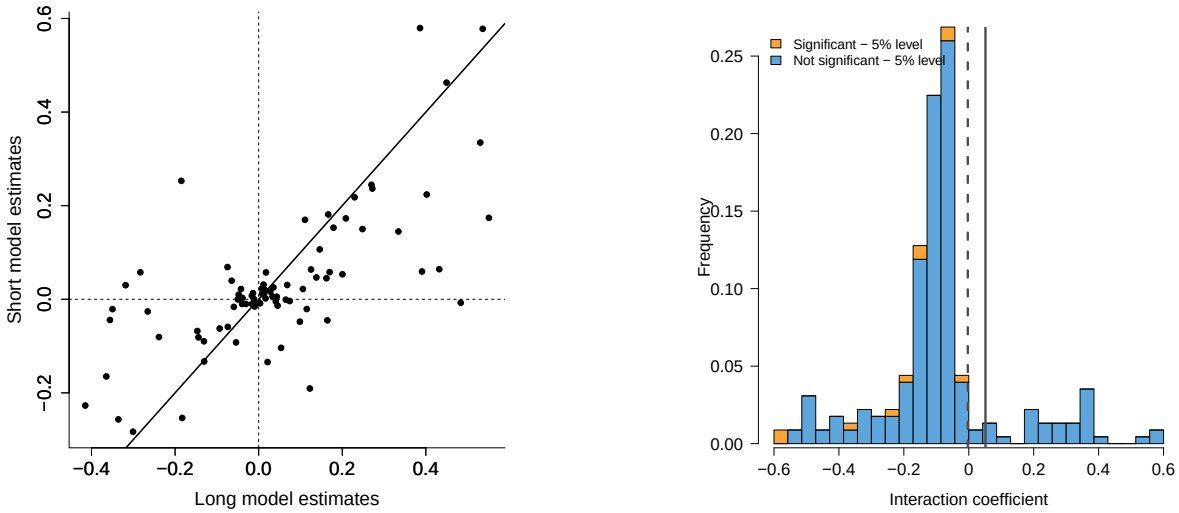


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

A.1.2 Ten most cited papers

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

(a) Main treatment estimates (b) Interaction



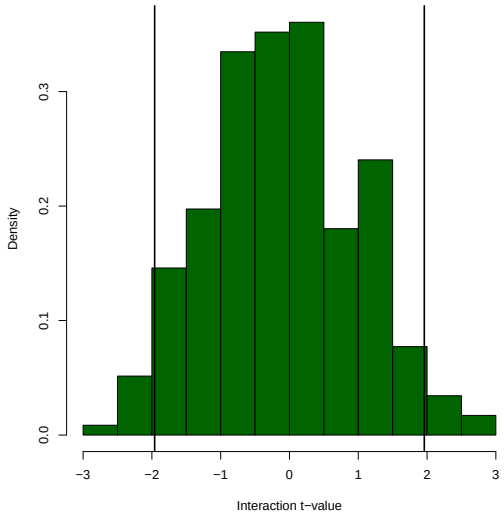
Note: Both figures show treatment estimates from the ten most cited papers with factorial designs and publicly available data that do not include the interactions in the original study. Figure A.2a shows how the main treatment estimates change across the short and the long model across studies (N=85 in this figure). The median main treatment estimate from the short model is 0.01σ, the median main treatment estimate from the long model is 0.01σ, the average absolute difference between the treatment estimates of the short and the long model is 0.05σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 131%, and 28% of treatment estimates change sign when they are estimated using the long instead of the short model. Figure A.2b shows the distribution of the interactions between the main treatments (N=266 in this figure). We trim the top and bottom 1% of the distribution. The median interaction is -0.00σ (dashed vertical line), the median absolute value of the interactions is 0.05σ (dashed vertical line), 5.6% of interactions are significant at the 10% level, 2.6% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.37.

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

Panel A: Significance at the 10% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	49	13	62
Significant	6	17	23
Total	55	30	85
Panel B: Significance at the 5% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	60	9	69
Significant	4	12	16
Total	64	21	85
Panel C: Significance at the 1% level			
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	73	3	76
Significant	1	8	9
Total	74	11	85

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

Figure A.3: Distribution of the t-value of interaction terms across studies

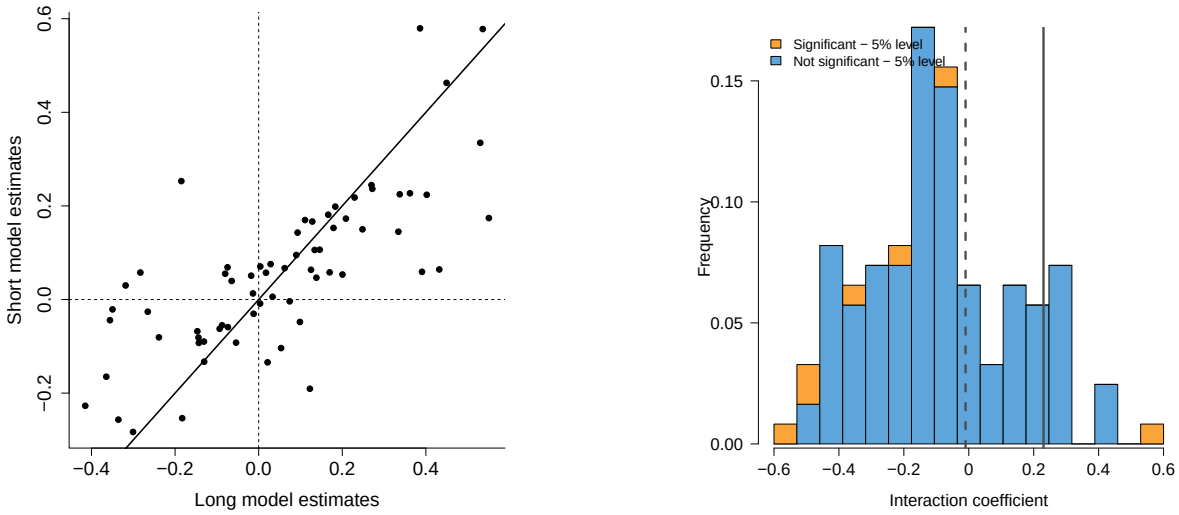


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

A.1.3 Policy experiments

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

(a) Main treatment estimates (b) Interaction



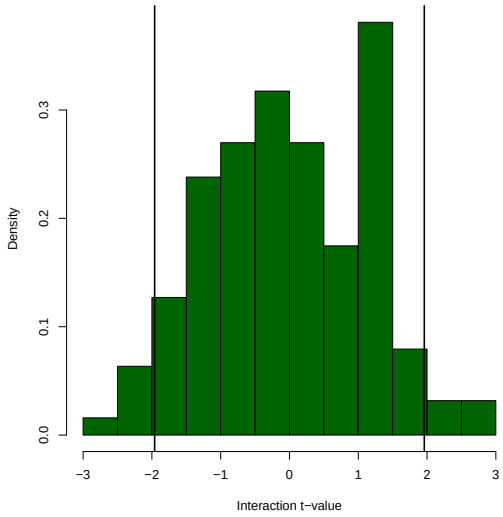
Note: Both figures show treatment estimates from the papers with factorial designs and publicly available data that do not include the interactions in the original study and do policy evaluation (N=67 in this figure). Figure A.4a shows how the main treatment estimates change across the short and the long model across studies. The median main treatment estimate from the short model is 0.06σ, the median main treatment estimate from the long model is 0.05σ, the average absolute difference between the treatment estimates of the short and the long model is 0.07σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 69%, and 21% of treatment estimates change sign when they are estimated using the long model instead of the short model. Figure A.4b shows the distribution of the interactions between the main treatments (N=126 in this figure). We trim the top and bottom 1% of the distribution. The median interaction is -0.01σ (dashed vertical line), the median absolute value of interactions is 0.23σ (solid vertical line), 6.3% of interactions are significant at the 10% level, 3.2% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 1.01.

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

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

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

Figure A.5: Distribution of the t-value of interaction terms across studies

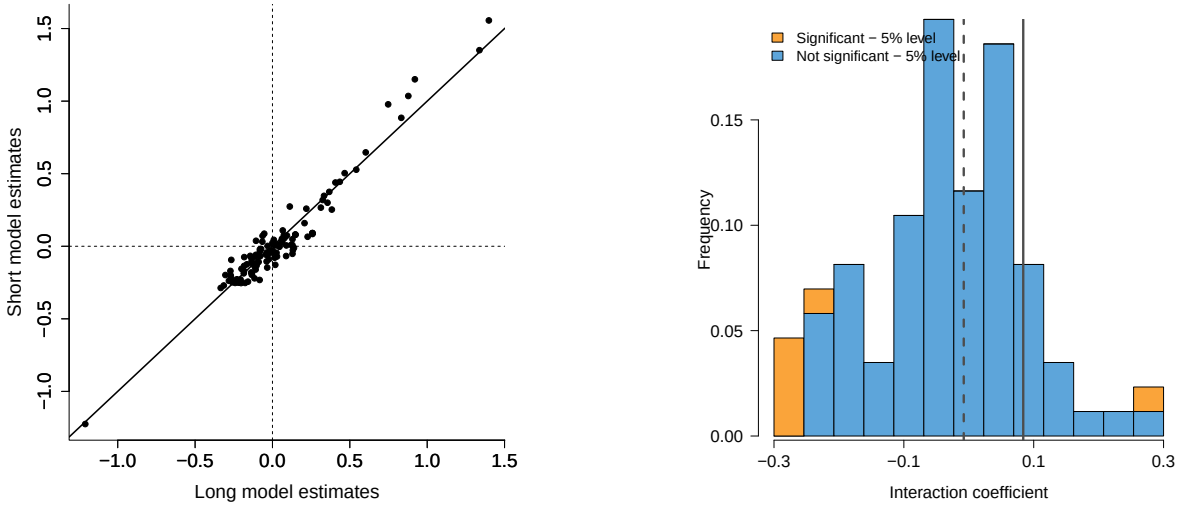


Note: If studies have factorial designs that cross-randomize more than two treatments only two-way interactions are included in this calculation. The vertical lines are at ± 1.96.

A.1.4 Studies with all interactions included

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

(a) Main treatment estimates (b) Interaction



Note: Both figures show treatment estimates from the papers with factorial designs and publicly available data that do not include the interaction in the original study and do policy evaluation. Figure A.6a shows how the main treatment estimates change across the short and the long model across studies (N=117 in this figure). The median main treatment estimate from the short model is -0.03σ, the median main treatment estimate from the long model is -0.02σ, the average absolute difference between the treatment estimates of the short and the long model is 0.05σ, the median absolute difference in percentage terms between the treatment estimates of the short and the long model is 37%, and 15% of treatment estimates change sign when they are estimated using the long or the short model. Figure A.6b shows the distribution of the interactions between the main treatments (N=104 in this figure). We trim the top and bottom 1% of the distribution. The median interaction is -0.01σ (dashed vertical line), the median absolute value of interactions is 0.08σ (solid vertical line), 4.5% of interactions are significant at the 10% level, 1.1% are significant at the 5% level, and 0.0% are significant at the 1% level, and the median relative absolute value of the interaction with respect to the main treatment effect is 0.52.

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

Panel A: Significance at the 10% level

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

Panel B: Significance at the 5% level

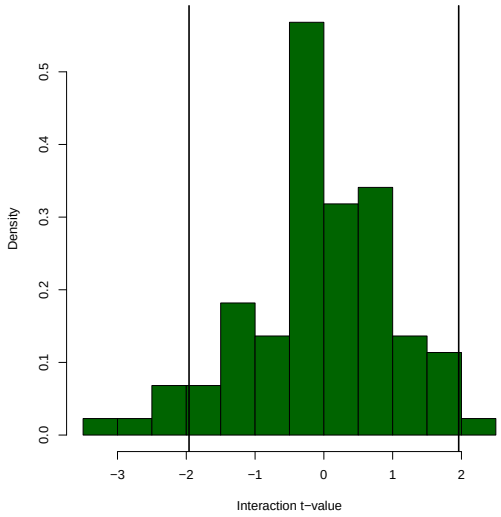
	Without interaction		
With interaction	Not significant	Significant	Total
Not significant	68	10	78
Significant	6	33	39
Total	74	43	117

Panel C: Significance at the 1% level

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

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

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



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

A.2 Derivation of expressions for the regression coefficients

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

Because the long regression model (1) is fully saturated, we have

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

Random assignment implies that, for $(t_1, t_2) \in \{0,1\} \times \{0,1\}$,

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

Thus, it follows that

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

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

Here we derive Equation (6). Equation (7) then follows from rearranging terms. The derivations of Equations (8) and (9) are similar and thus omitted.

For the short regression model (2), independence of T_1 and T_2 implies that

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

Consider

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

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

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

Thus, we have

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

A.3 Power to detect interactions in the long model

Under the assumptions in Section 2.4, the standard errors of the long model are

$$SE(\hat{\beta}_1) = \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \quad \text{and} \quad SE(\hat{\beta}_{12}) = \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4}}.$$

To achieve power κ, the true interaction effect needs to satisfy (e.g., Duflo et al., 2007)

$$\beta_{12} > \left(\Phi^{-1}(\kappa) + \Phi^{-1}(1 - \alpha/2)\right) SE(\hat{\beta}_{12}) = MDE_{\beta_{12}}.$$

where α is the size of the test. Here MDE stands for minimum detectable effect size. Similarly, to achieve power κ for detecting the main effect, it must satisfy

$$\beta_1 > \left(\Phi^{-1}(\kappa) + \Phi^{-1}(1 - \alpha/2)\right) SE(\hat{\beta}_1) = MDE_{\beta_1}.$$

We can relate the MDEs to the overall sample size required for detecting interactions,

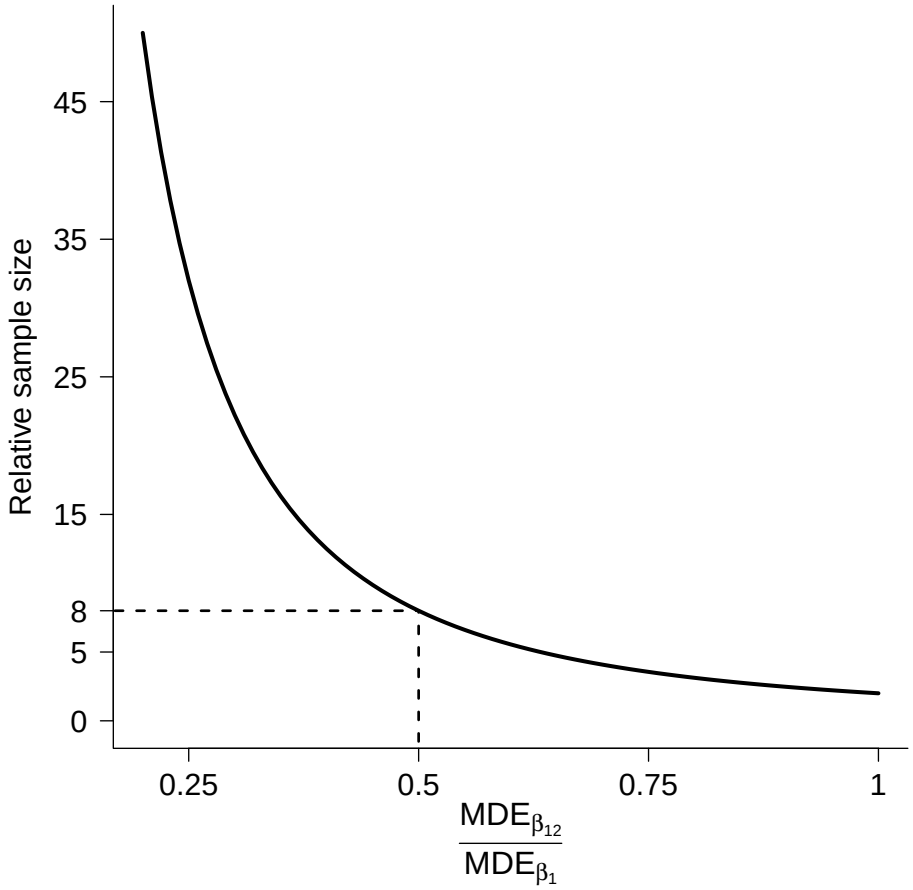
N_I, and main effects, N_M, respectively. To illustrate, suppose that the overall sample size is equally distributed across all four cells (the power-maximizing design for detecting interactions). In this case, the standard errors are $SE\left(\hat{\beta}_1\right) = \sigma\sqrt{8/N_M}$ and $SE\left(\hat{\beta}_{12}\right) = \sigma\sqrt{16/N_I}$, such that

$$\frac{MDE_{\beta_1}}{MDE_{\beta_{12}}} = \frac{\sigma\sqrt{\frac{8}{N_M}}}{\sigma\sqrt{\frac{16}{N_I}}} \qquad \text{and} \qquad 2\left(\frac{MDE_{\beta_1}}{MDE_{\beta_{12}}}\right)^2 = \frac{N_I}{N_M}.$$

Suppose that the MDE for the interaction effect is half the MDE for the main effect. Then the relative sample size needed to be adequately powered (N_I/N_M) is 8. That is, we need eight times the sample size to detect an interaction effect that is half the size of the main effect.[1] Even if the MDE for the interaction is the same as the MDE for the main effect, one would need twice the sample size to detect the interaction effect than to detect the main effect. Figure A.8 illustrates the general relationship between N_I/N_M and $MDE_{\beta_{12}}/MDE_{\beta_1}$.

[1] Alternatively, one can compare $MDE_{\beta_{12}}$ to the MDE based on the short model, $MDE_{\beta_1^s}$, as in Gelman (2018). Because $SE(\hat{\beta}_1^s) = \sigma\sqrt{4/N}$, the required sample size for detecting an interaction is 16 times larger than for detecting main effects based on the short model.

Figure A.8: Relative sample size

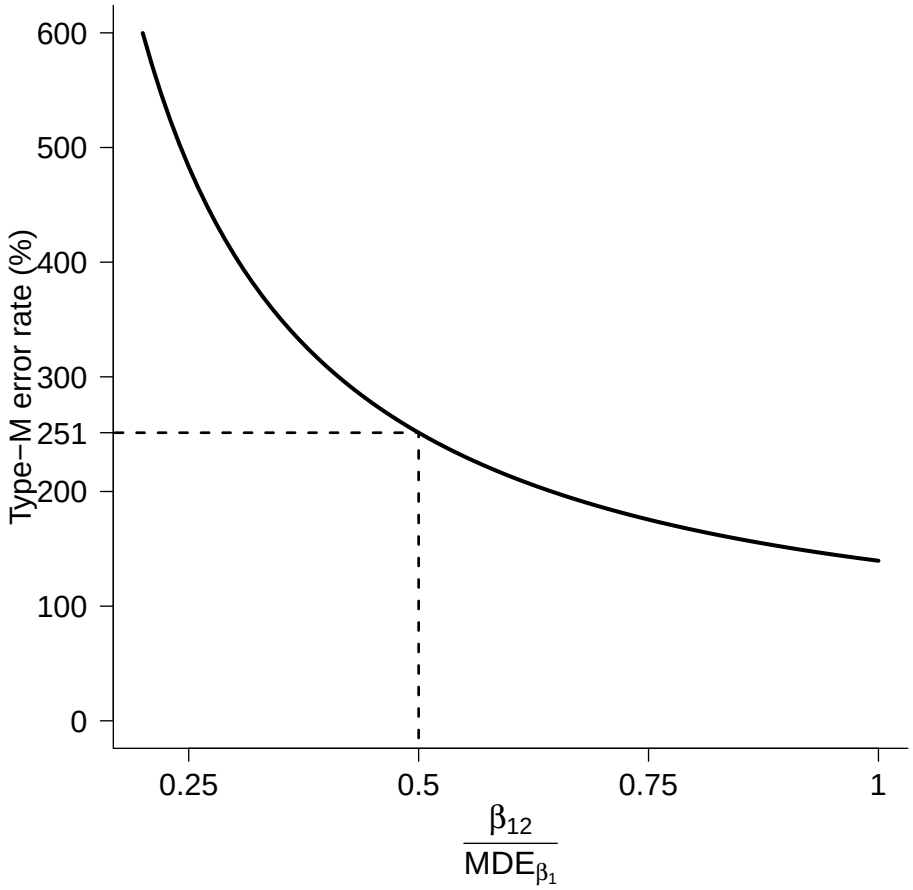


Note: For We assume the sample is divided equally among the four cells in Table 1. This figure plots the relative sample size $\left(\frac{N_I}{N_M}\right)$ *as a function of the relative MDEs* $\left(\frac{MDE_{\beta_{12}}}{MDE_{\beta_1}}\right)$.

Figure A.9 shows the Type-M error for different values of the interaction (relative to the MDE of the main effect, which determines the sample size).[2] We use the closed form formula provided by Lu et al. (2019) for the Type-M error. In the figure, we assume the MDE for the main effect is 0.2σ (or equivalently, a sample of 1,570 equally divided among the four cells, assuming size is $\alpha = 0.05$ and power is $\kappa = 0.8$).

[2] A related problem with under-powered studies is the *Type S error rate*, which is the probability that conditional on being significant, the estimate of the interaction in a hypothetical replication study based on the same design as the original study has an incorrect sign (see p.643 in Gelman & Carlin, 2014).

Figure A.9: Type-M error



Note: We assume the sample is divided equally among the four cells in Table 1. This figure plots the Type-M error for different values of the interaction (relative to the MDE of the main effect, which determines the sample size). We use the closed form formula provided by Lu et al. (2019) for the Type-M error. We assume that size is $\alpha = 0.05$, power is $\kappa = 0.8$, and the MDE for the main effect is 0.2σ (i.e., a sample of 1,570 equally divided among the four cells).

A.4 Detailed description of the econometric methods

A.4.1 The EMW approach

To describe Elliott et al. (2015)'s nearly optimal test, note that under standard conditions, the t-statistics are approximately normally distributed in large samples

$$\begin{pmatrix} \hat{t}_1 \\ \hat{t}_{12} \end{pmatrix} \overset{a}{\sim} N\left(\begin{pmatrix} t_1 \\ t_{12} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{13}$$

where $\hat{t}_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$, $\hat{t}_{12} = \frac{\hat{\beta}_{12}}{SE(\hat{\beta}_{12})}$, $t_1 = \frac{\beta_1}{SE(\hat{\beta}_1)}$, $t_{12} = \frac{\beta_{12}}{SE(\hat{\beta}_{12})}$, and $\rho = Cov(\hat{t}_1, \hat{t}_{12})$. Define $\hat{t} = (\hat{t}_1, \hat{t}_{12})$ and $t = (t_1, t_{12})$. In practice, we replace the unknown $SE(\hat{\beta}_1)$, $SE(\hat{\beta}_{12})$, and $Cov(\hat{t}_1, \hat{t}_{12})$ with heteroskedasticity robust estimators, which are consistent under

weak conditions.

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

$$H_0: \ t_1 = 0, \ \ t_{12} \in \mathbb{R} \quad \text{against} \quad H_1: \ t_1 \neq 0, \ \ t_{12} = 0. \tag{14}$$

A common approach to construct powerful tests for problems with composite hypotheses is to choose tests based on their weighted average power. In particular, we seek a powerful test for "H_0: the density of $\hat{t}$ is $f_t, t_1 = 0, t_{12} \in \mathbb{R}$" against the simple alternative "$H_{1,F}$: the density of $\hat{t}$ is $\int f_t dF(t)$", where the weighting function F is chosen by the researcher. Following Elliott et al. (2015), we choose F so that it assigns equal mass to 2 and -2. To obtain the best test, one needs to find a LFD, Λ^{LF}, such that the size α Neyman-Pearson test of $H_{0,\Lambda^{LF}}$ against $H_{1,F}$ is also a size α test of H_0 against $H_{1,F}$, where $H_{0,\Lambda}$: the density of $\hat{t}$ is $\int f_t d\Lambda(t)$ (Lehmann & Romano, 2005; Elliott et al., 2015).

Since it is generally difficult to analytically determine and computationally approximate Λ^{LF}, Elliott et al. (2015) suggest to instead focus on an approximate LFD, Λ^{ALF}, which yields a nearly optimal test for H_0 against $H_{1,F}$. The resulting test is then just a Neyman-Pearson test based on Λ^{ALF}.[3]

A.4.2 The AKK approach

To describe Armstrong et al. (2020)'s approach, we write model (11) in vector form as

$$\mathbf{Y} = \beta_1 \mathbf{T}_1 + \beta_{12} \mathbf{T}_{12} + \varepsilon. \tag{15}$$

Suppose that $\mathbf{X} = (\mathbf{T}_1, \mathbf{T}_{12})$ is fixed and $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_N)$, where σ^2 is known.[4] The algorithm we describe below accommodates non-Gaussian and heteroskedastic errors. A linear estimator of β_1 can be written as $\hat{\beta}_1 = a'\mathbf{Y}$, for some a that can depend on $\mathbf{X}$. Given parameters (β_1, β_{12}), the bias of $\hat{\beta}_1$ is $a'(\beta_1 \mathbf{T}_1 + \beta_{12} \mathbf{T}_{12}) - \beta_1$. The "worst case" bias of $\hat{\beta}_1$ is

$$\overline{\text{bias}} = \sup_{\beta_1 \in \mathbb{R}, \beta_{12} \in [-C, C]} a'(\beta_1 \mathbf{T}_1 + \beta_{12} \mathbf{T}_{12}) - \beta_1. \tag{16}$$

The standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1) = \sigma\sqrt{a'a}$, does not depend on (β_1, β_{12}).

The t-ratio $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$ is normally distributed, $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(b, 1)$, where $|b| \leq \frac{\overline{\text{bias}}}{SE(\hat{\beta}_1)}$.

[3] 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.

[4] If $\mathbf{X}$ is random, the procedure remains valid, as it is valid conditional on $\mathbf{X}$.

Thus, a two-sided $(1-\alpha)$ confidence interval centered at $\hat{\beta}_1$ can be constructed as

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

where $cv_\alpha(x)$ is the $(1-\alpha)$ quantile of a $|N(x,1)|$ distribution. The length of the confidence interval (17) is increasing in $\overline{\text{bias}}$ and $SE(\hat{\beta}_1)$. Thus, to construct optimal confidence intervals, a is chosen to solve this bias variance trade-off.[5] Armstrong et al. (2020) show that this problem can be solved using a regularized regression of $\mathbf{T}_1$ on $\mathbf{T}_{12}$.

We use Algorithm 3.1 in Section 3 of Armstrong et al. (2020), which accommodates heteroskedastic and non-Gaussian errors.[6] To describe the algorithm, let π^*_λ denote the solution to the following penalized regression problem

$$\min_\pi \|\mathbf{T}_1 - \pi \mathbf{T}_{12}\|_2^2 + \lambda|\pi|. \tag{18}$$

The algorithm has three steps.

1. Compute initial estimates of the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N$ from the long regression model and obtain an initial variance estimator $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_i^2$.

2. Compute the solution path $\{\pi^*_\lambda\}_{\lambda>0}$ for the regularized regression (18), indexed by λ. For each λ, compute $\hat{\beta}_{1,\lambda}$ as

$$\hat{\beta}_{1,\lambda} = \frac{(\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12})'\mathbf{Y}}{(\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12})'\mathbf{T}_1} \tag{19}$$

   and obtain $\overline{\text{bias}}_\lambda$ and SE_λ as

$$\overline{\text{bias}}_\lambda = \frac{C}{|\pi_\lambda|} \frac{(\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12})'\mathbf{T}_{12}\pi^*_\lambda}{(\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12})'\mathbf{T}_1} \quad \text{and} \quad SE^2_\lambda = \frac{\hat{\sigma}^2 \|\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12}\|_2^2}{\left[(\mathbf{T}_1 - \pi^*_\lambda \mathbf{T}_{12})'\mathbf{T}_1\right]^2}. \tag{20}$$

3. Choose $\lambda^* = \arg\min_\lambda cv_\alpha \left(\frac{\overline{\text{bias}}_\lambda}{SE_\lambda} \right) SE_\lambda$ and compute robust standard errors $\widehat{SE_{\text{r},\lambda^*}} = \sqrt{\sum_{i=1}^N a^2_{\lambda^*,i}\hat{\varepsilon}_i^2}$, where $a_{\lambda^*} = \frac{(\mathbf{T}_1 - \pi^*_{\lambda^*}\mathbf{T}_{12})}{(\mathbf{T}_1 - \pi^*_{\lambda^*}\mathbf{T}_{12})'\mathbf{T}_1}$. Return the optimal $(1-\alpha)$ confidence

[5] 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 can be different.

[6] The implementation of the optimal confidence intervals with potentially heteroskedastic and non-Gaussian errors mimics the common practice of applying OLS in conjunction with heteroskedasticity robust standard errors, rather than weighted least squares; see Remark 3.2 in Armstrong et al. (2020) for a discussion.

interval

$$\hat{\beta}_{1,\lambda^*} \pm cv_\alpha \left(\frac{\widehat{\text{bias}}_{\lambda^*}}{\widehat{SE}_{\text{r},\lambda^*}} \right) \widehat{SE}_{\text{r},\lambda^*}. \tag{21}$$

A.4.3 The IMS approach

For a given $\beta_{12} \in [C_1, C_2]$, the population regression coefficient from a regression of $Y - \beta_{12} T_{12}$ on $X = (1, T_1, T_2)'$ is

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

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

$$\beta_t \in [\beta_t(C_2), \beta_t(C_1)] = \left[\beta_t^l, \beta_t^u\right].$$

The lower bound β_t^l can be estimated from an OLS regression of $Y - C_2 T_{12}$ on X. Similarly, the upper bound β_t^u can be obtained from an OLS regression of $Y - C_1 T_{12}$ on X. Under standard conditions, the OLS estimators $\hat{\beta}_t^l$ and $\hat{\beta}_t^u$ are asymptotically normal and the asymptotic variances $Avar\left(\hat{\beta}_t^l\right)$ and $Avar\left(\hat{\beta}_t^u\right)$ can be estimated consistently. We can therefore apply the approach of Imbens & Manski (2004) and Stoye (2009) to construct confidence intervals for β_t:[7]

$$CI_{1-\alpha} = \left[\hat{\beta}_t^l - c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^l\right)}{N}},\ \hat{\beta}_t^u + c_{IM} \cdot \sqrt{\frac{\widehat{Avar}\left(\hat{\beta}_t^u\right)}{N}} \right], \tag{22}$$

where the critical value c_{IM} solves

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

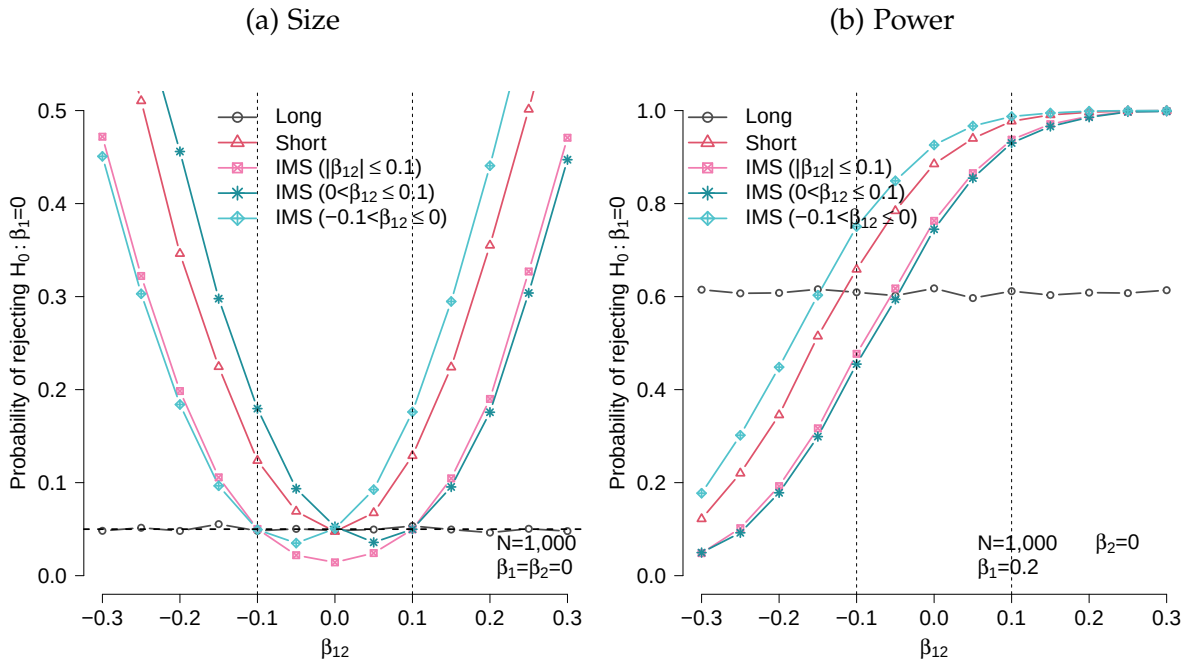
By Imbens & Manski (2004) and Stoye (2009), $CI_{1-\alpha}$ is a valid confidence interval for β_t.

In the running example in the main text we imposed $C_1 = -0.1$ and $C_2 = 0.1$.

[7] By construction, $P(\hat{\beta}_t^u \geq \hat{\beta}_t^l) = 1$ so that Lemma 3 in Stoye (2009) ensures that the conditions in Imbens & Manski (2004) hold.

Figure A.10 shows size and power of the IMS approach for $C_1 = 0$ and $C_2 = 0.1$ and for $C_1 = -0.1$ and $C_2 = 0$.

Figure A.10: Two sided and one-sided restrictions on β_{12} under IMS



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for Figures 6a and 6b is $\alpha = 0.05$. *IMS refers to the Imbens & Manski (2004) and Stoye (2009) approach for constructing valid confidence intervals under prior knowledge about the magnitude of* β_{12}. *The dashed vertical lines are placed at* $\beta_{12} = -0.1$ *and* $\beta_{12} = 0.1$.

A.5 Econometric details for the design-based solution

A.5.1 Power improvements

Consider a factorial design with an empty interaction cell as in Section 4.4 (see Table A.5) and the following population regression model

$$Y = \beta_0^* + \beta_1^* T_1 + \beta_2^* T_2 + \varepsilon^*. \tag{23}$$

Let $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ denote the OLS estimators of β_1^* and β_2^*. If T_1 and T_2 are randomly assigned, $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are consistent for the respective main effects (see Appendix A.5.2).[8]

To illustrate the power implications of leaving the interaction cell empty, consider an experiment where the researcher cares equally about power to detect an effect of T_1 and T_2, and thus assigns the same sample size to both treatments: $N_2^* = N_3^* = N_T^*$. To illustrate, we focus on β_1^*. The variance of $\hat{\beta}_1^*$ is given by $Var\left(\hat{\beta}_1^*\right) = \sigma^2 \frac{N - N_T^*}{(N - 2N_T^*)N_T^*}$.

[8]Note in this case T_1 and T_2 are not independent of each other because of the negative correlation between the probability of being assigned to T_1 and T_2.

$Var\left(\hat{\beta}_1^*\right)$ is minimized when $N_T^* = \frac{N}{2}\left(2-\sqrt{2}\right)$ and we assume that the experiment is designed in this manner.[9] A comparison to the variance of the estimator based on the long model, $\hat{\beta}_1$, shows that $Var\left(\hat{\beta}_1^*\right) \leq Var\left(\hat{\beta}_1\right)$.[10] Thus, leaving the interaction cell empty yields power improvements for testing hypotheses about the main effects relative to long model t-tests.

Table A.5: Leaving the interaction cell empty

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

A.5.2 Consistency of the OLS estimators based on model (23)

Here we show that when the interaction cell is empty and T_1 and T_2 are randomly assigned, the OLS estimators based on the regression model (23) are consistent for the main effects.

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

Multiplying out yields the following expressions for β_1^*:

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

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

$$\beta_1^* = \frac{-p_1 p_2 E(Y) + p_1 p_2(1 - p_2)E(Y \mid T_1 = 1) + p_1 p_2^2 E(Y \mid T_2 = 1)}{-p_1^2 p_2 - p_1 p_2^2 + p_1 p_2}. \tag{24}$$

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

$$E(Y) = E(Y \mid T_1 = 1, T_2 = 0)p_1 + E(Y \mid T_1 = 0, T_2 = 0)(1 - p_1 - p_2) + E(Y \mid T_1 = 0, T_2 = 1)p_2. \tag{25}$$

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

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

Combining (24) and (25) and simplifying yields:

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

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

A.6 Bonferroni-correction with consistent model selection

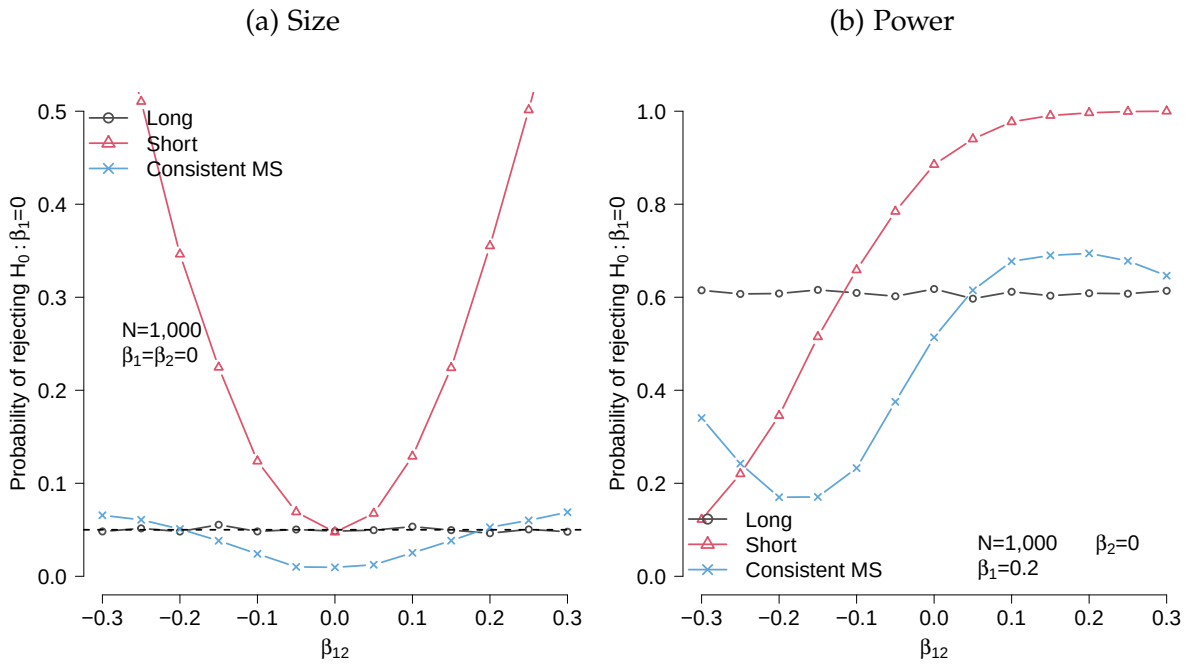
In Section 4.2, we discussed a nearly optimal test that yields power improvements over the long model t-test near a priori likely values of the interaction. Here we discuss an alternative to the nearly optimal test: the Bonferroni approach of McCloskey (2017, 2020).

To achieve size control in the presence of model selection, one could employ tests based on the largest critical value across all possible values of the interaction effect β_{12}. However, this so-called least favorable approach is known to be very conservative due to its worst case nature. McCloskey (2017, 2020) suggests a procedure that improves upon the least favorable approach and asymptotically controls size. The basic insight of this approach is that one can construct an asymptotically valid confidence interval for β_{12}. As a consequence, one can search for the largest critical value over the values of β_{12} in the confidence interval rather than over the entire real line as in the least favorable approach. The uncertainty about the nuisance parameter (β_{12}) and the test statistic can be accounted for using a Bonferroni-correction.

McCloskey (2017, 2020) considers both conservative and consistent model selection. Under conservative model selection, one uses a fixed threshold to select the model irrespective of the sample size. An example is the model selection algorithm in Section 2.5 where one employs a 5% t-test in the first step, irrespective of the sample size. Under consistent model selection, the model selection threshold is allowed to grow with the sample size. We explored both approaches and found that consistent model selection leads to more powerful tests in our context. We therefore only report results for consistent model selection. Specifically, we implement the adjusted Bonferroni critical values outlined in Section 3.2 of McCloskey (2017) and in Section 5 of McCloskey (2020).[11]

[11]Specifically, we use the algorithm "Bonf-Adj Post-Sel" outlined in both papers. We employ consistent model selection using the BIC criteria. We use $\beta = 0.5$ which results in $\overline{a} = 0.45$, as suggested by McCloskey (2017). To speed our simulations, we use the true OLS standard errors (as opposed to the estimated ones).

Figure A.11: McCloskey (2017, 2020)'s consistent model selection exhibits small size distortions and yields power gains over running the full model for positive values of β_{12}



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for Figures 5a and 5b is $\alpha = 0.05$. Consistent MS refers to McCloskey (2017, 2020)'s consistent model selection.

Figure A.11 reports the results of applying McCloskey (2017, 2020)'s Bonferroni-style correction to our running example. It shows that consistent model selection with state-of-the-art Bonferroni adjustments leads to local power improvements relative to the long model for a short range of positive values of the interaction effect β_{12}. However, unlike the nearly optimal test discussed in Section 4.2, researchers cannot choose where those power gains occur.

As expected, these power improvements come at the cost of much lower power for other values of β_{12}. While the Bonferroni-correction asymptotically controls size for all values of the interaction, we find some small size distortions in our simulations. Appendix A.7.7 provides a more comprehensive assessment of the performance by plotting power curves for different values of β_1.

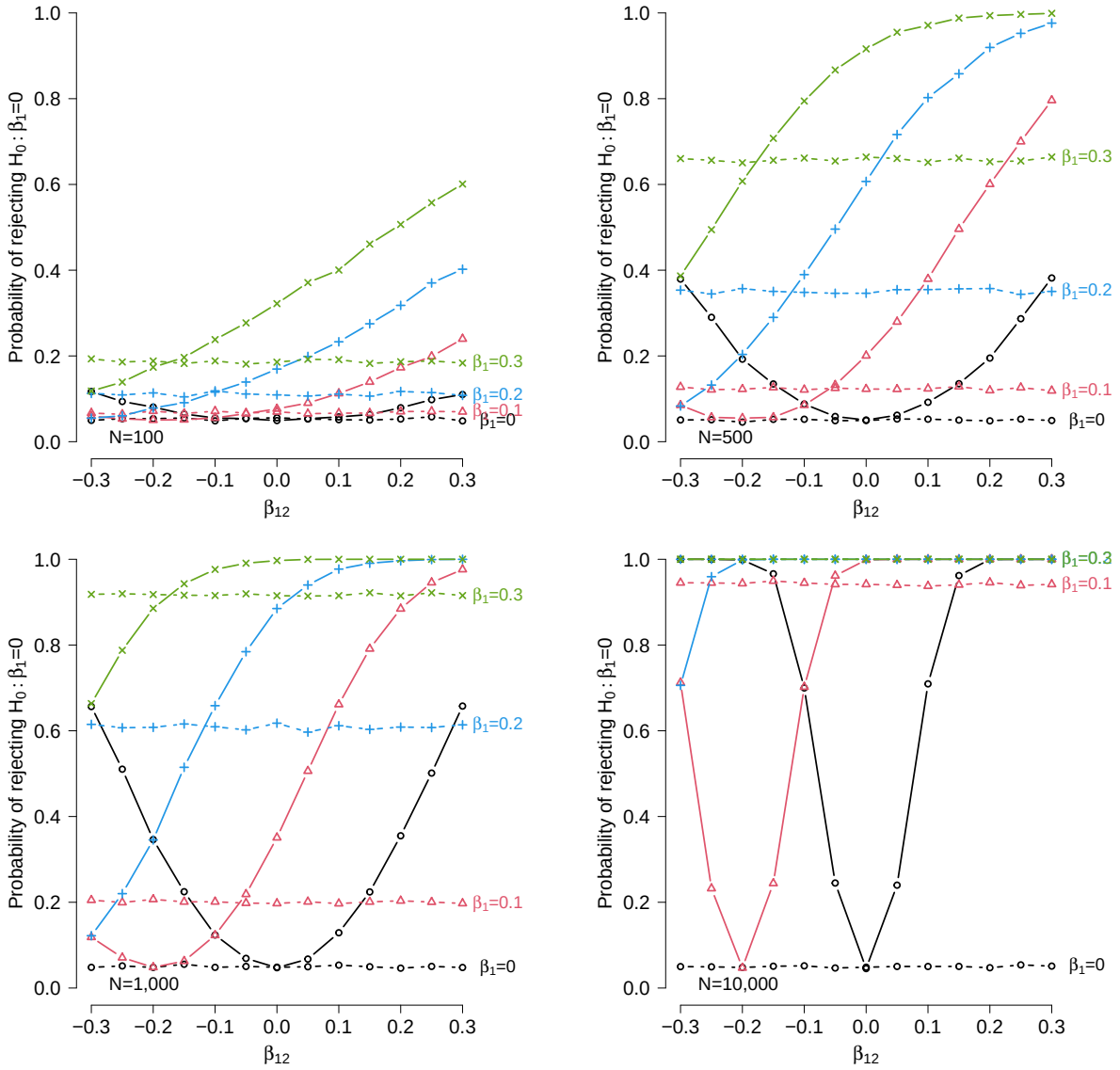
A.7 Additional figures and tables

Table A.6: Articles published in top-5 journals between 2007 and 2017

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

A.7.1 Ignoring the interaction

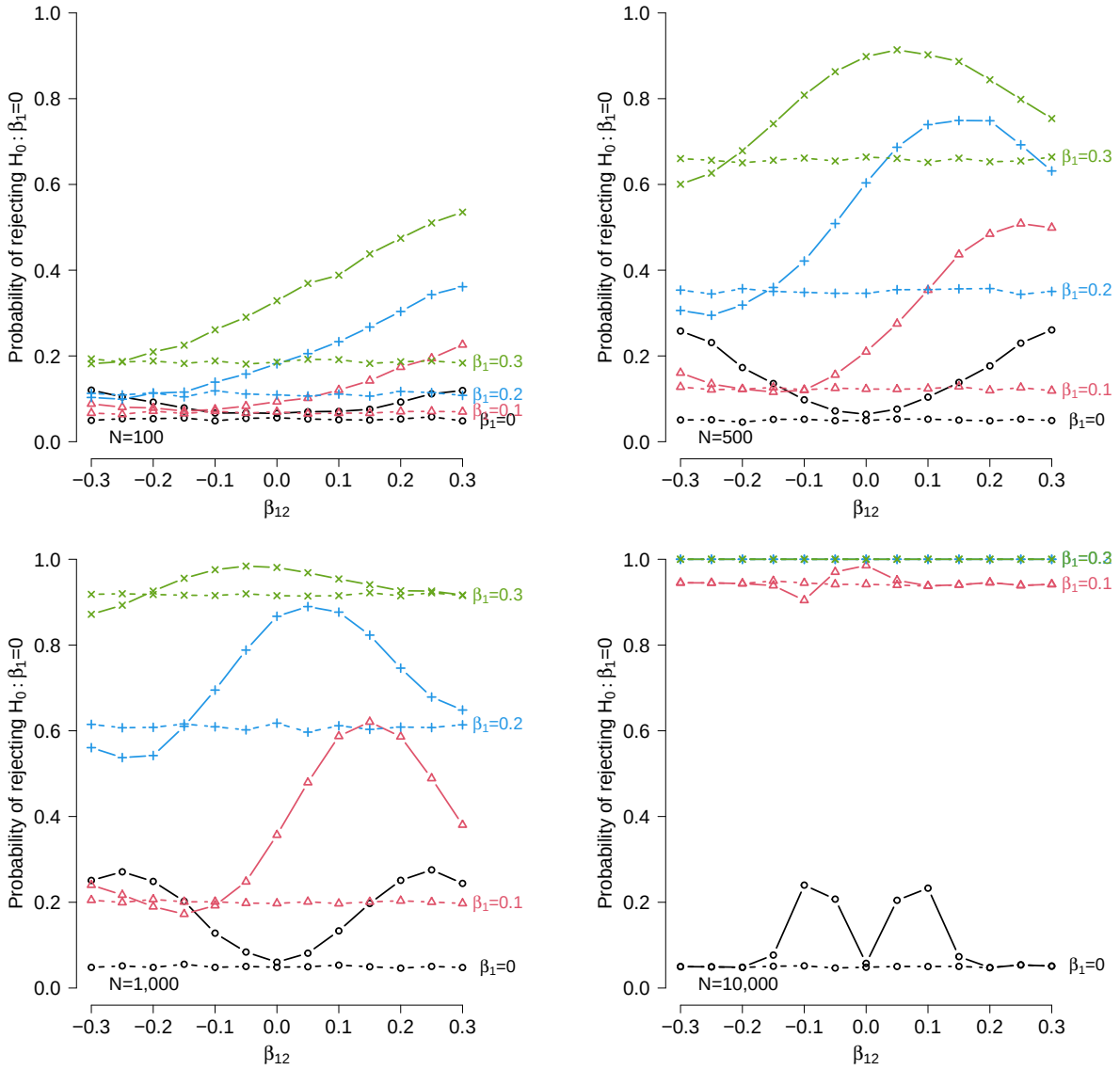
Figure A.12: Long and short model: Power curves



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

A.7.2 Model selection (pre-testing)

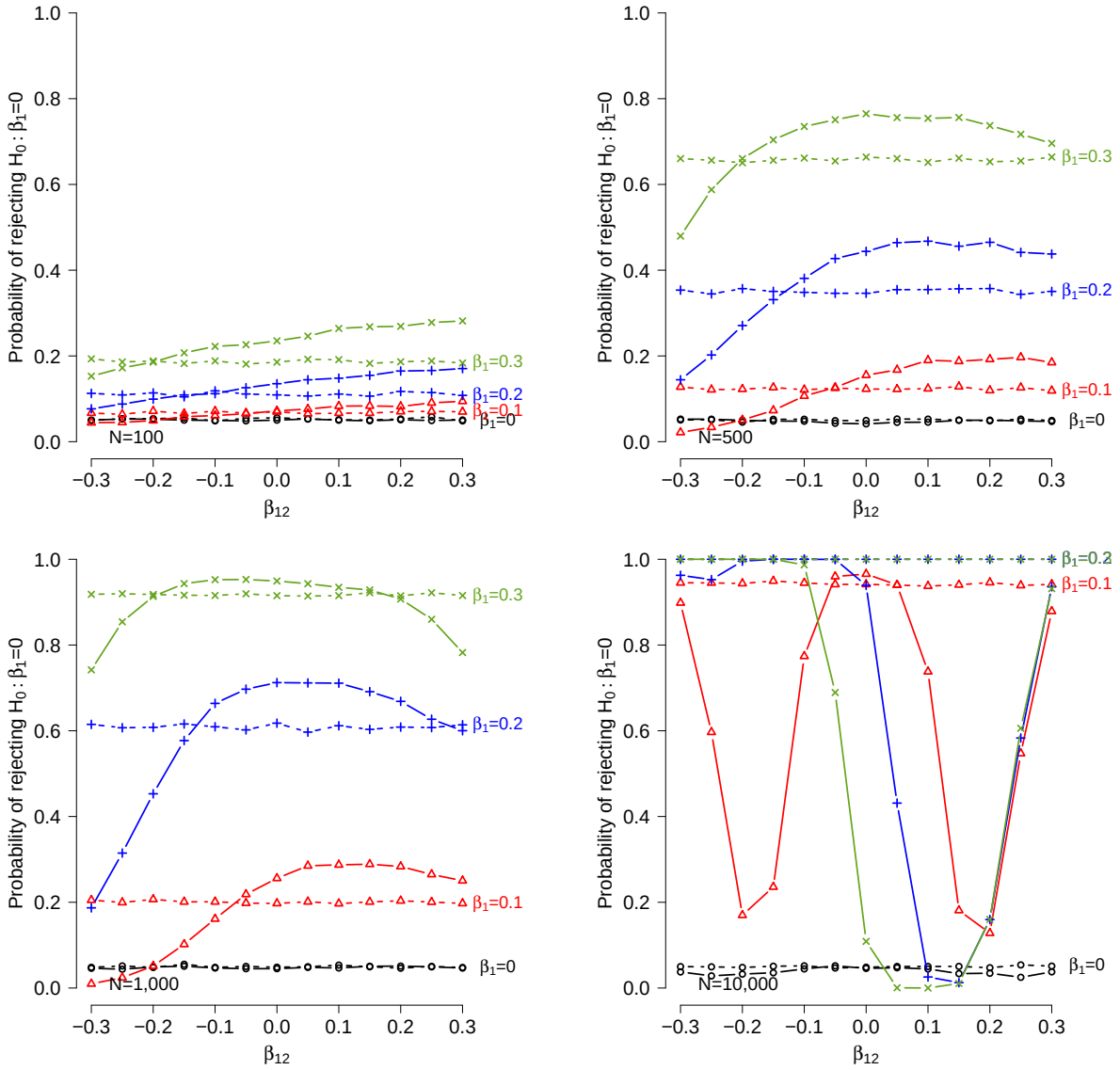
Figure A.13: Long model and model selection: Power curves



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

A.7.3 Elliott et al. (2015)'s nearly optimal test

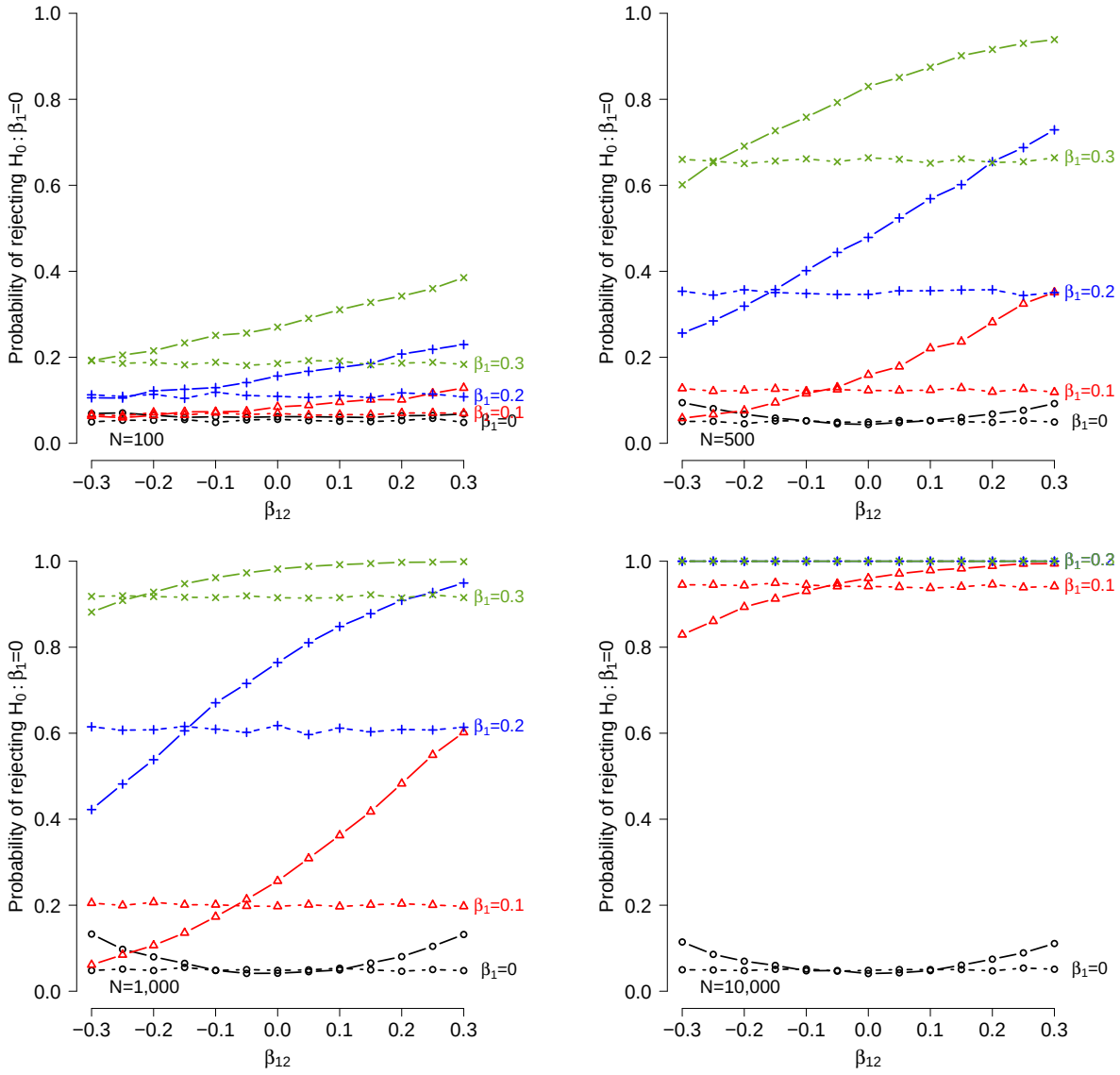
Figure A.14: Long model and Elliott et al. (2015)'s nearly optimal test: Power curves



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

A.7.4 Restrictions on the magnitude of β_{12}: Armstrong et al. (2020)

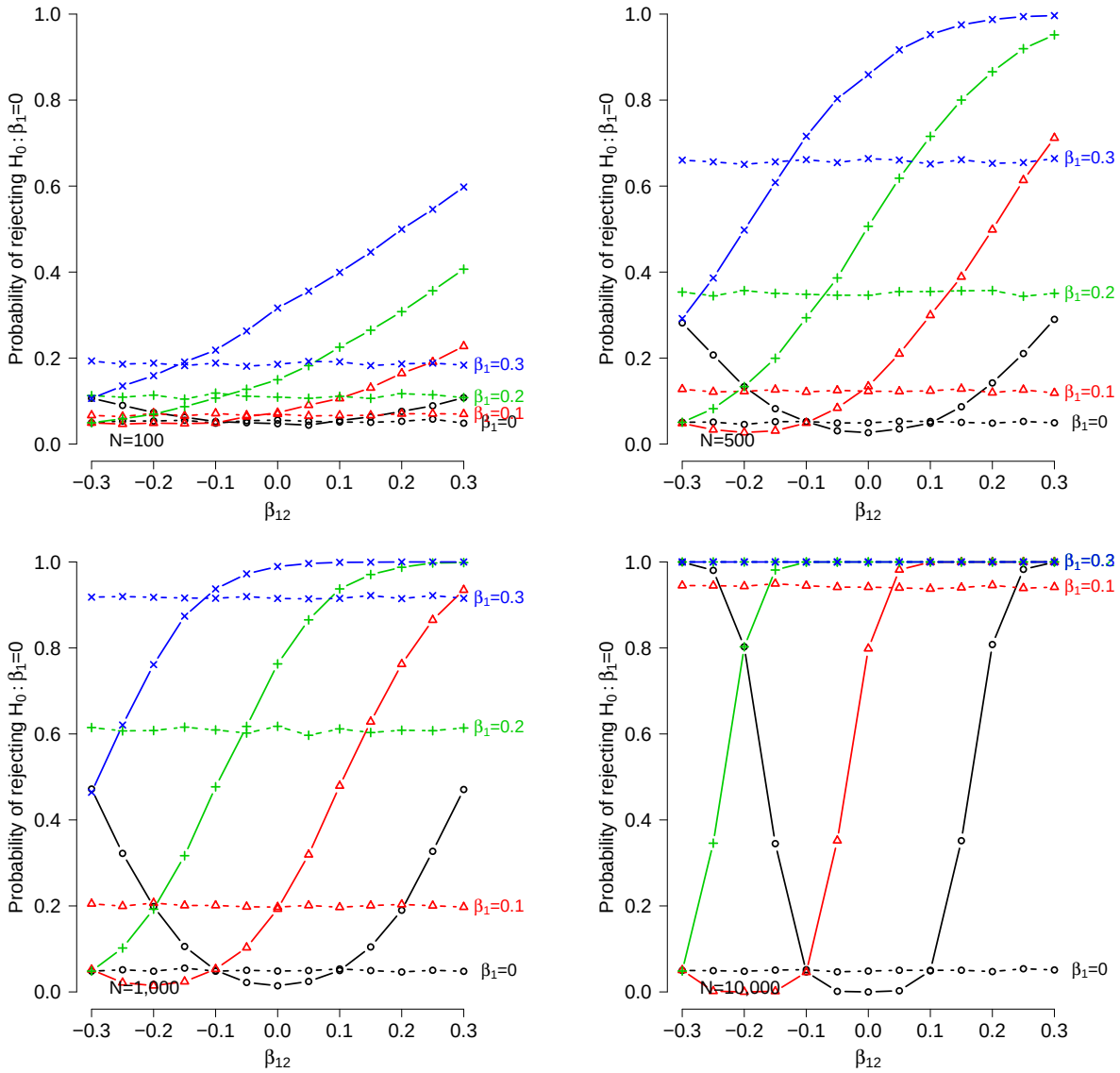
Figure A.15: Long model and Armstrong et al. (2020)'s approach: Power curves



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. *In each figure, dashed lines show the power for the long model, while solid lines show power for Armstrong et al. (2020)'s approach based on restrictions on the magnitude of* β_{12}.

A.7.5 Restrictions on the magnitude of β_{12}: Imbens & Manski (2004) and Stoye (2009)

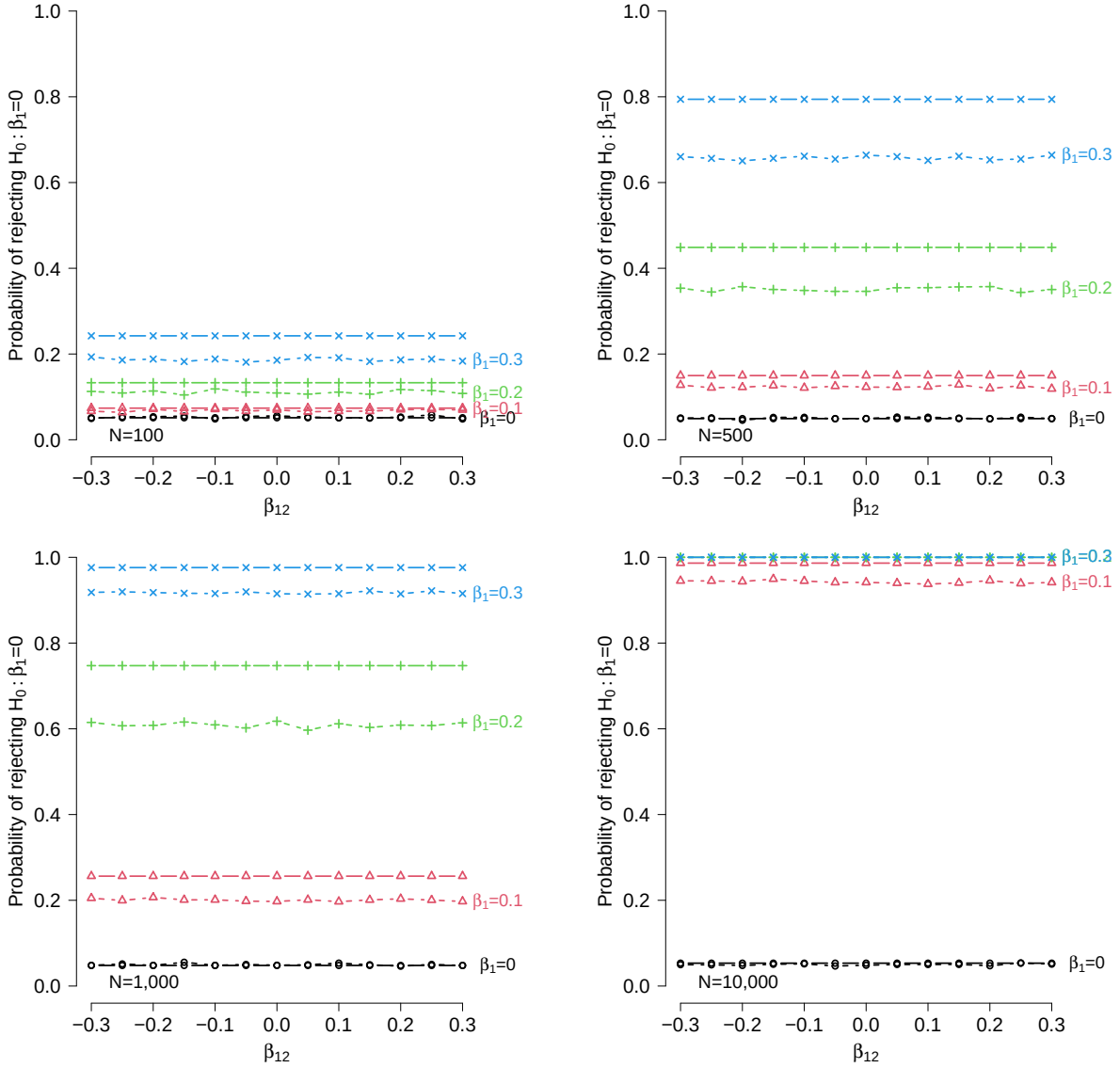
Figure A.16: Long model and Imbens & Manski (2004) and Stoye (2009)'s approach: Power curves



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. In each figure, dashed lines show the power for the long model, while solid lines show power for Imbens & Manski (2004) and Stoye (2009)'s approach based on restrictions on the magnitude of β_{12}.

A.7.6 Leaving the interaction cell empty

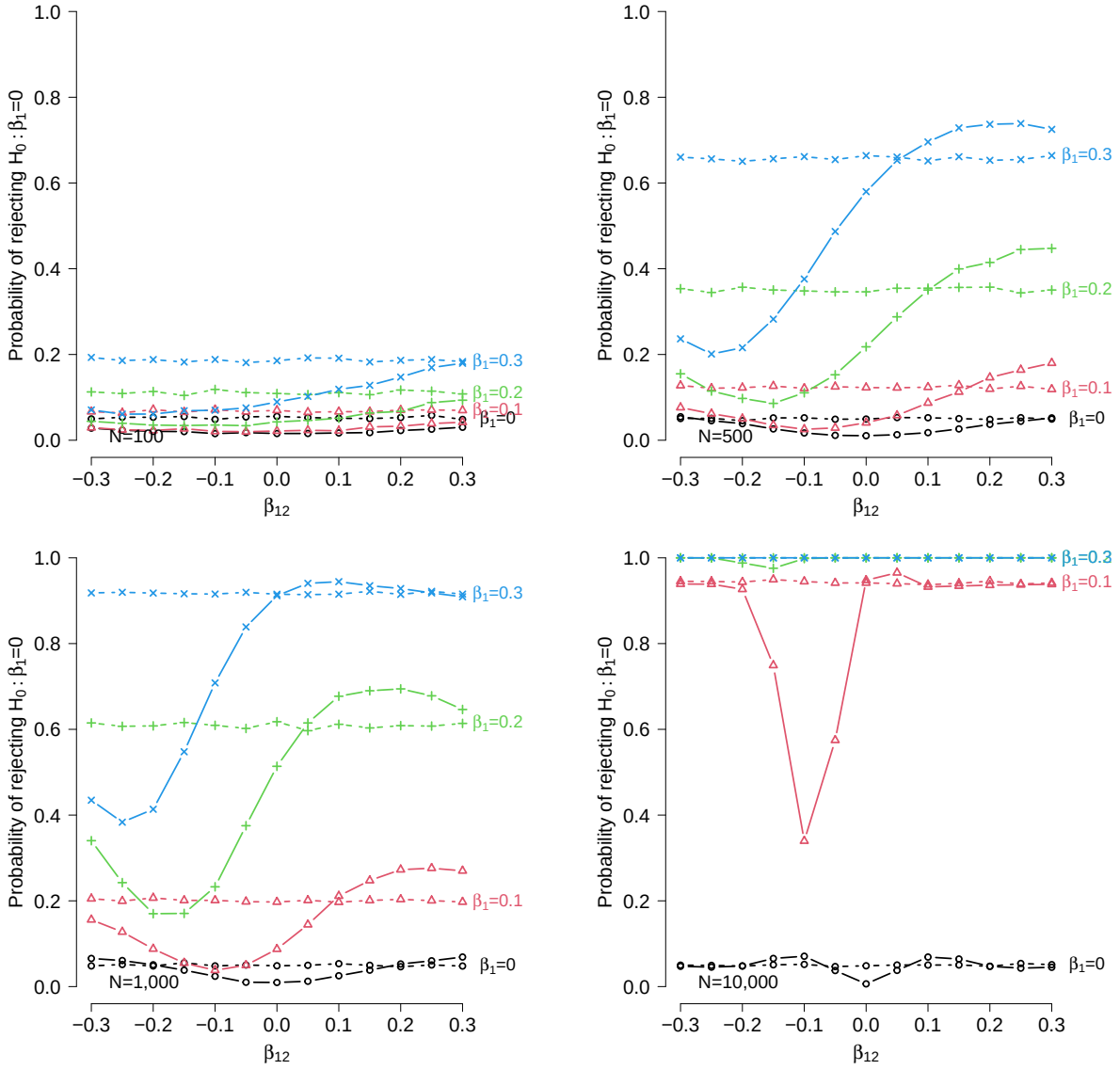
Figure A.17: Long model and leaving the interaction cell empty: Power curves



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$. *In each figure, dashed lines show the power for the long model, while solid lines show power a design with the same sample size but leaving the interaction cell empty.*

A.7.7 McCloskey (2017, 2020)'s consistent model selection with Bonferroni-type correction

Figure A.18: Long model and McCloskey (2017, 2020)'s consistent model selection with Bonferroni-type correction: Power curves

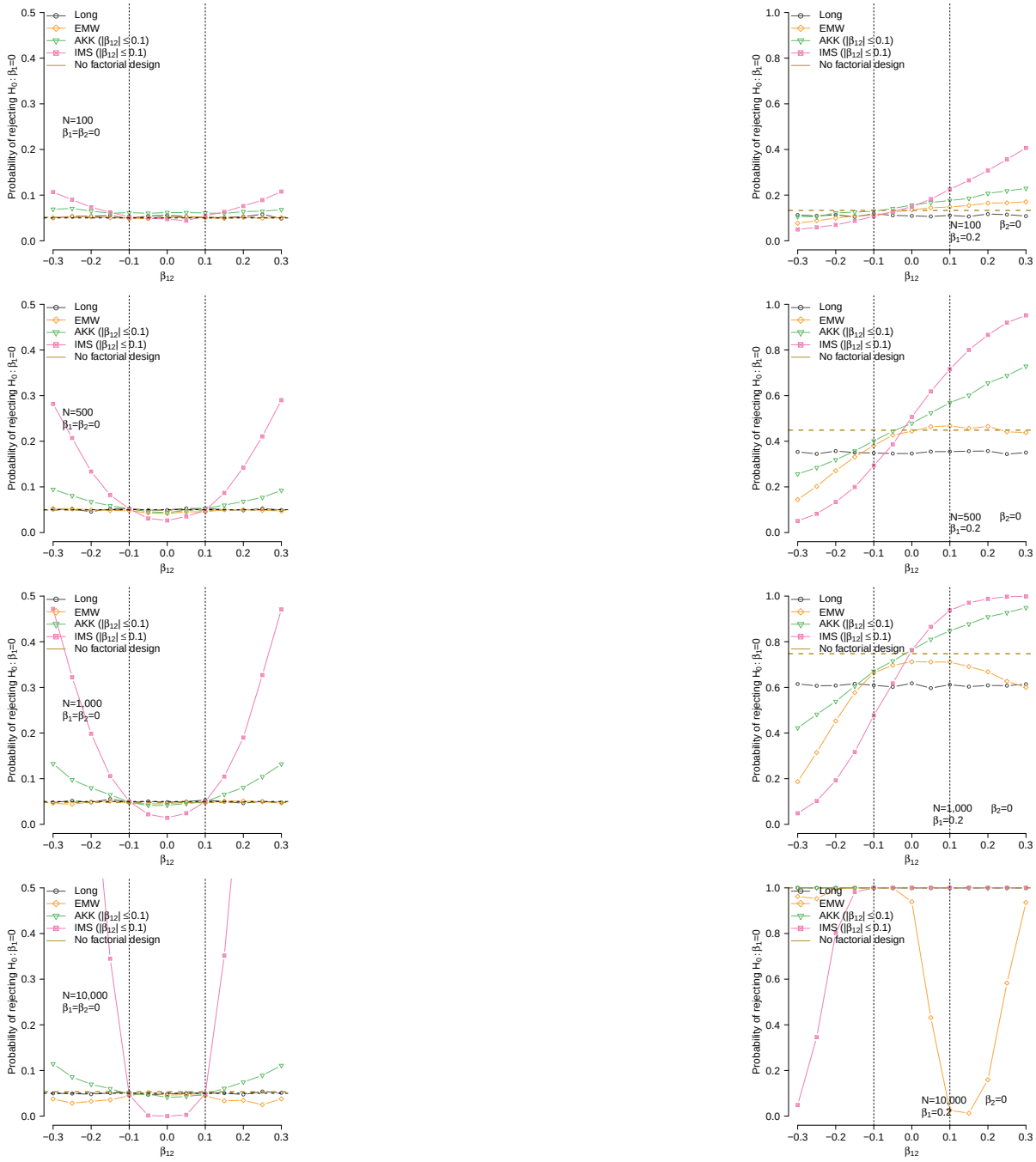


Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size across all figures is $\alpha = 0.05$*. In each figure, dashed lines show the power for the long model, while solid lines show power for McCloskey (2017, 2020)'s consistent model selection with Bonferroni-type correction.*

A.7.8 Comparison across methods

Figure A.19: No factorial design: Size and power

(a) Size (b) Power



Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. $N_T^* = 0.29N$ *and* $N_1^* = 0.42N$*. The size across all figures is* $\alpha = 0.05$*.*

B Short description of each paper with a factorial design considered in "Factorial designs, model selection, and (incorrect) inference in randomized experiments"

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

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

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

TABLE 1
NUMBER OF VILLAGES IN EACH TREATMENT CATEGORY

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

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

Note: Table 1 from Olken (2007).

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

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

The paper does present the interactions *after* the main tables, which are estimated using the short model. Explicitly, "Panel B of Table IV compares the Balsakhi and the CAL effects and examines their interactions in year 2 (2002-2003) when they were implemented at the same time using a stratified design. When the two programs are considered in isolation, the CAL has a larger effect on math test scores than the Balsakhi Program (although this difference is not significant) and a smaller effect on overall test scores (although, again, the difference is not significant). The programs appear to have no interaction with each other: the coefficients on the interaction on the math and overall test score are negative and insignificant." However, the paper does not mention that the estimates based on the short model must be interpreted as weighted averages of treatment effects with respect to the different counterfactuals defined by the other treatments in the experiment.

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

TABLE I
SAMPLE DESIGN AND TIME LINE

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

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

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

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

The evaluation featured a factorial design with three treatments: Extra contract teacher; school-based management; and tracking (i.e., splitting classes by ability). Figure B.3 taken from Duflo et al. (2008) working paper has details of the experimental design. The published version of the paper does not mention the school-based management treatment. The long model is not presented in any table in the paper, nor in the appendix. The paper does not mention that the estimates based on the short model must be interpreted as weighted averages of treatment effects with respect to the different counterfactuals defined by the other treatments in the experiment. We re-estimate the results of Table IV (Panel A, Column 1) in Duflo et al. (2011) using the long model.[1].

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

Figure 1
Experimental Design: The Extra-Teacher Project

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

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

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

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

Kleven et al. (2011) analyze a tax enforcement field experiment in Denmark. The experiment features a factorial design with two independent treatments. The first is a random audit and the second is threat-of-audit letters. The data are not available online. The main tables in the paper use the short model to estimate treatment effects. The paper does not mention that the estimates based on the short model must be interpreted as weighted averages of treatment effects with respect to the different counterfactuals defined by the other treatments in the experiment. After the main tables, Table VI analyzes the effects of one treatment (information letters) conditional on the other treatment (audit), from which they conclude that "letter effects are roughly the same in the 0% and 100% audit groups."

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

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

B.6 Agricultural Decisions after Relaxing Credit and Risk Constraints

Karlan et al. (2014) conduct several field experiments in Ghana. Farmers were randomly assigned to receive cash grants, a rainfall index insurance, or a combination of the two. The main tables in the paper (Table IV – Table VII) estimate the fully saturated long model. The data are not available online.

B.7 What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment

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

B.8 The Demand for, and Impact of, Learning HIV Status

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

B.9 The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya

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

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

Figure B.4: Factorial design in Haushofer & Shapiro (2016)

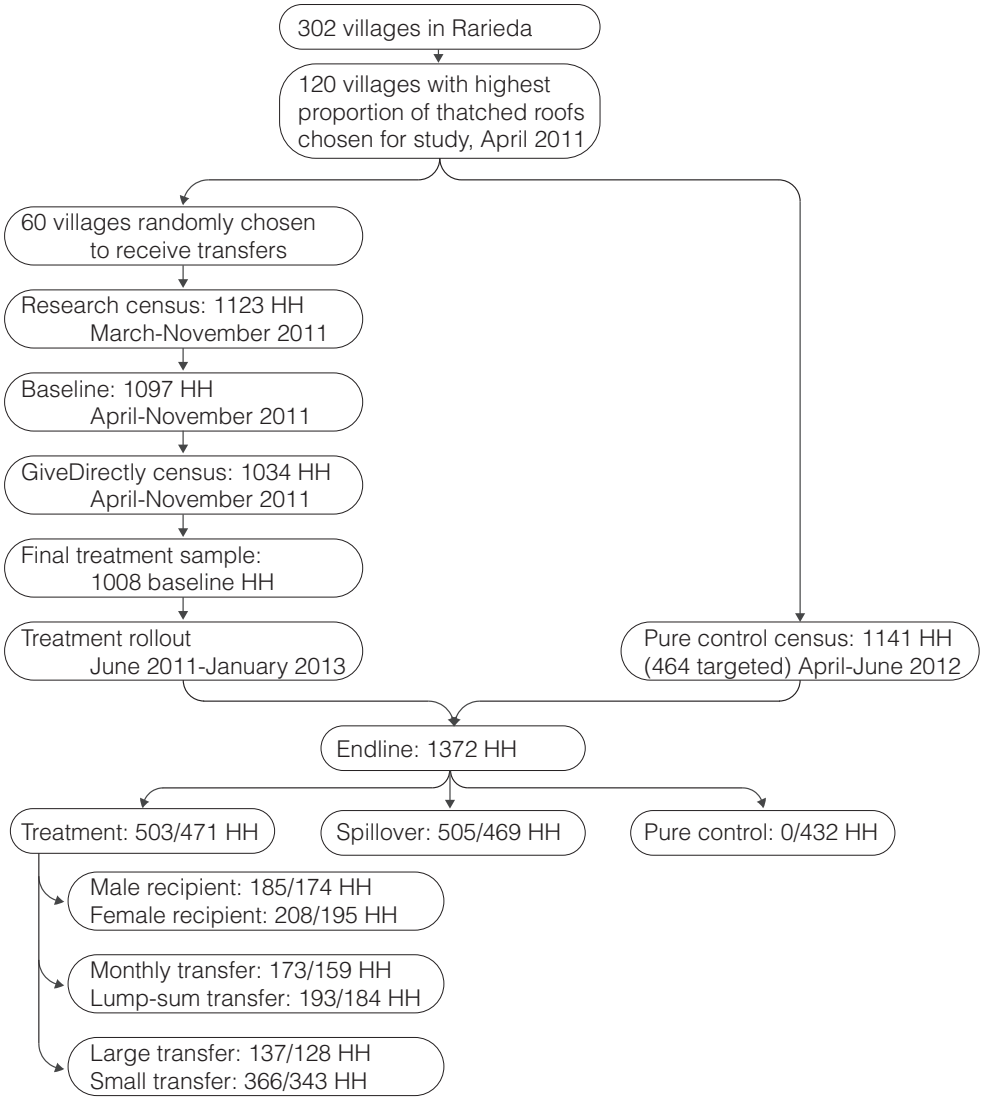


FIGURE I

Timeline of Study

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

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

B.10 Targeting the Poor: Evidence from a Field Experiment in Indonesia

Alatas et al. (2012) analyze an experiment in Indonesia, in which villages are randomly assigned to different targeting methods to distribute a cash transfer program. In some villages targeting is done using a proxy-means test, in some targeting is done by the community, and in some is a hybrid of both. In "community" and "hybrid" villages the treatments had several variations: In some villages, the meetings took place during the day, in others at night. In some, the "elite" of the village took the decision, in some, it was the whole community. In some, the 10 poorest households were primed by the meeting facilitator, in some, there was no priming. Explicitly, the paper states "We designed several subtreatments in order to test three hypotheses about why the results from the community process might differ from those that resulted from the PMT treatment: elite capture, community effort, and within-community heterogeneity in preferences." Figure B.5 taken from Alatas et al. (2012) has details of the experimental design. However, the paper does not mention that the treatment effects in the main tables (e.g., Tables 3 and 4) are the weighted average over the subtreatments. Explicitly, the paper states "the PMT treatment is the omitted category, so β_1 and β_2 are interpretable as the impact of the community and the hybrid treatments relative to the PMT treatment". After the main results, Tables 7 explores the "elite" subtreatment. We re-estimate the results in Table 3 (Column 1) including all possible interactions.

Figure B.5: Factorial design in Alatas et al. (2012)

TABLE 1—RANDOMIZATION DESIGN

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

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

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

B.11 Credit Elasticities in Less-Developed Economies: Implications for Microfinance

Karlan & Zinman (2008) analyze an experiment in South Africa in which a lender sent out direct mail offers to over 50,000 former clients. The letters had a randomly assigned offer interest rate and in some cases a randomly assigned, nonbinding example maturity (four, six, or twelve months). In addition, each client was assigned a randomly selected a "contract rate" that was weakly less than the offer rate received by mail and revealed only after the borrower had accepted the solicitation and applied for a loan. We do not study the re-randomization of the interest rate.[3] However, the paper does not mention that the estimates in the main tables (e.g., Table 3 looking at the interest rate) should be interpreted as weighted averages of treatment effects with respect to different counterfactuals. None of the tables in the main paper or the

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

appendix estimate the long model. We re-estimate the results in Table 3 (Column 1) and Table 8 (Column 1) including the interaction between the interest rate and the example maturity.

B.12 Education, HIV, and Early Fertility: Experimental Evidence from Kenya

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

B.13 Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving

Andreoni et al. (2017) analyze a field experiment with two interventions where they placed people soliciting donations for The Salvation Army Red Kettle Campaign. They have a 2×2 design where "Solicitation occurred in two modes: only bell ringing or bell ringing with a verbal request...In the opportunity conditions, solicitors rang the bell as usual but did not speak or attempt eye contact, except to thank those who gave, as per Red Kettle custom. The ask condition was the same as the opportunity condition except that solicitors attempted eye contact with each passerby and said, "Hi, how are you? Merry Christmas. Please give today." The other dimension is whether we had solicitors at only door 1 or at both doors 1 and 2." They use the long model throughout the paper. We re-estimate Table 2.

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

B.14 Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies

Jakiela & Ozier (2015) analyze an experiment to measure the impacts of social pressure to share income with kin and neighbors in rural Kenyan villages. To do this they assign participants to one of six treatments in a 2 x 3 design. Explicitly, "Within the experiment, players were randomly assigned to one of six treatments. First, players were allocated either the smaller endowment of 80 shillings or the larger endowment of 180 shillings....Every player was also assigned to either the private treatment or one of two public information treatments, the public treatment or the price treatment". They use the long model throughout the paper. We re-estimate Table 2 in the paper in the form of a long regression with interactions.

B.15 Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment

Eriksson & Rooth (2014) study whether long-term unemployment spells matter for employers hiring decisions using a field experiment. The experimental design varies several applicant characteristics. Explicitly, "[t]he applicants were randomly assigned a number of attributes which typically are included in job applications and are expected to be important for the probability of being invited to a job interview. These attributes include contemporary and past spells of unemployment, work experience, education, gender, ethnicity, and some other characteristics." Each application was randomly assigned different characteristics using a factorial design. The following characteristics (and their possible values) were randomized: 1) Unemployment duration (takes value 0, 3, 6, or 9), 2) unemployed before employment (takes values 0 or 1), 3) unemployed between jobs (takes values 0 or 1), 4) work experience (takes values 1, 2, 3, 4 or 5), 5) number of employers (takes values 0 or 1), 6) ethnicity/gender (the applicant randomized to be native male, native female or ethnic minority male), 7) having more education than required (takes values 0 or 1), 8) work experience during the summer breaks (takes values 0 or 1), 9) visiting US high school (takes values 0 or 1), 10) Personality trait I - agency (takes values 0 or 1), 11) Personality trait II - communion (takes values 0 or 1), and 12) leisure activities (randomized to have one of seven different leisure activities or none). As the authors explicitly state: "The typical approach in field experiments using the correspondence testing methodology is to vary only one characteristic in the applications, e.g., ethnicity or gender of the applicant (cf. Riach and Rich 2002; Carlsson and Rooth 2007). However, in our experiment, we used a more general approach by randomly varying several characteristics. This allows us to measure the labor market return of different skills and attributes (cf.

Bertrand and Mullainathan 2004; Rooth 2011)." The paper does not mention that the estimates in the main tables (e.g., Table 6) should be interpreted as a weighted average of treatment effects relative to different counterfactuals, nor does it estimate the full model in the paper or in the appendix. We re-estimate Table 6: Column 1 using the long model including all possible two-way interactions, but there are higher-order interactions implied by the factorial design.

B.16 Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market

Allcott & Taubinsky (2015) report on two experiments, both of which have a 2×2 designs. Figure B.6 — taken from the published version of the paper — shows the details of the first experiment randomization design. Explicitly "Each consumer was randomly assigned to Treatment or Control, and within Treatment to a matrix of four subtreatments. These group assignments determined which two information screens the consumer would receive.... the "Positive" subtreatment included information about the cost savings from CFLs, while the "Balanced" subtreatment included information about cost savings and the CFL's negative attributes. The right column in the matrix of subtreatments is the Endline-only treatment, in which consumers skipped the baseline choices and began directly with the information provision. Except when specified, we pool these four subtreatments together and refer to them as the "Treatment" group; we show in Section IIIE that the effects of these four subtreatments are not statistically distinguishable."

Figure B.6: Factorial design in Allcott & Taubinsky (2015)

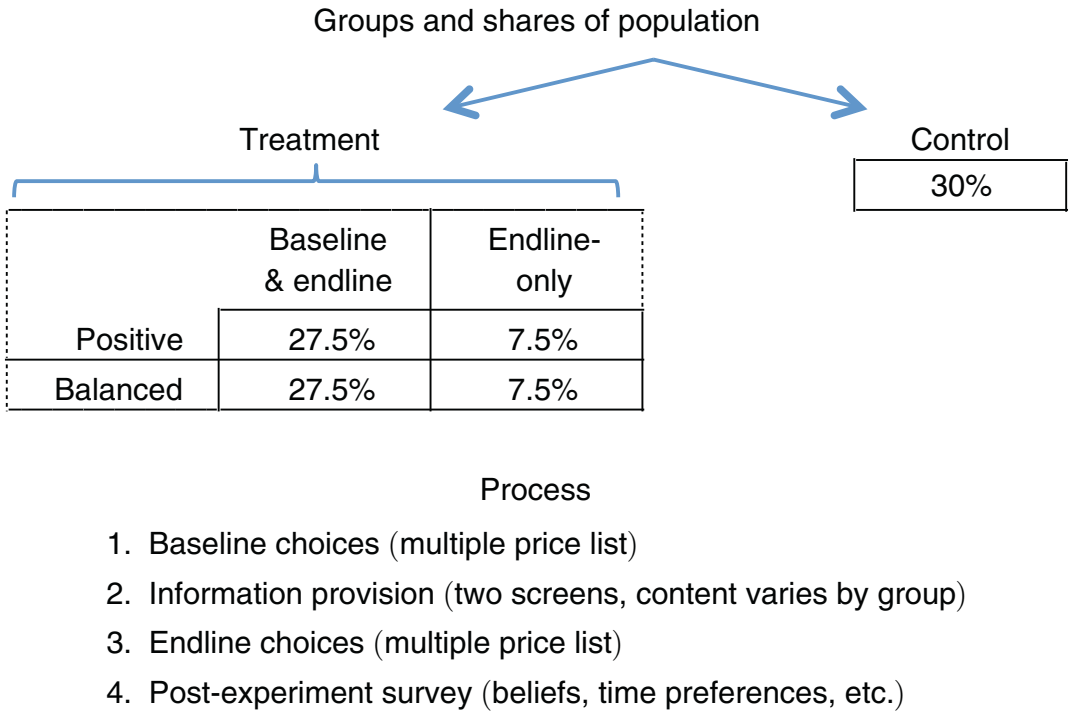


FIGURE 1. TESS EXPERIMENTAL DESIGN

Note: Figure 1 from Allcott & Taubinsky (2015).

The data for this experiment are available online. For this experiment, the paper does not mention that the estimates in the main tables (e.g., Table 1) should be interpreted as weighted averages of treatment effects with respect to different counterfactuals. Moreover, the text suggests they performed model selection.

The second experiment "Customers who consented were given a brief survey via iPad... The iPad randomized customers into information Treatment and Control groups with equal probability. For the Treatment group, the iPad would display the annual energy costs for CFLs versus incandescents, given the customer's estimated daily usage, desired wattage, and desired number of bulbs. The treatment screen also displayed the energy costs and total user costs (energy plus bulbs) for CFLs versus incandescents over the 8,000-hour rated life of a CFL.... At the end of the survey and potential informational intervention, the RAs gave customers a coupon in appreciation for their time. The iPad randomized respondents into either the Standard Coupon group, which received a coupon for 10 percent off all lightbulbs purchased, or the Rebate Coupon group, which received the same 10 percent coupon plus a second coupon valid for 30 percent off all CFLs purchased. Thus, the Rebate Coupon

group had an additional 20 percent discount on all CFLs." For this experiment, the paper presents both the short and the long model (see Table 5), but focuses on the former. The data for this experiment is not publicly available.

B.17 Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions

Flory et al. (2014) analyze two experiments. The first experiment uses a 2×6×2 design in which the employment advertisement, compensation scheme, and application procedure vary. In the first dimension, ads for the job either "had masculine connotations or... a general ad that has removed those masculine connotations". In the second dimension, the experiment "randomized job-seekers who expressed interest in the position into one of six different treatments". In the third dimension, the experiment varied the application procedure. Explicitly, "The application questionnaires were randomized at the city level. In eight cities, job-seekers had to fill out a long questionnaire with four interview questions, while in the other eight cities the questionnaire was short and contained only one question." In the paper they do not use the city-level randomization on the length of the instrument, and neither do we since it does not appear in the data. The paper does not mention that the estimates based on the short model must be interpreted as weighted averages of treatment effects with respect to different counterfactuals. The second experiment does not have a factorial design.

The estimation compares male and female applications for the different employment advertisements in the different compensation schemes. We re-estimate a linear probability model (the paper uses logit models) for the likelihood of applying for a job using the long regression interacting all the treatments (the closest analog would be Table 7 in the paper), separately for males and females (as in the paper).

B.18 Shrouded Attributes and Information Suppression: Evidence from the Field

Brown et al. (2010) use several experiments to study the revenue effect of varying the level and disclosure of shipping charges in online auctions. The main tables (e.g., Table II) estimate the fully saturated long model. The data are not available online.

B.19 Voting to Tell Others

DellaVigna et al. (2016) analyze the results from a field experiment designed to estimate a model of voting "because others will ask". To do this, they use a factorial

design with four dimensions. First, households were randomized into five flyer treatments with equal weights, where the information received in a flyer varied across treatments. Then, they randomized the duration of the survey (5 minutes or 10 minutes). The third dimension randomized how the surveyors described the survey to the respondent. The fourth dimension randomized the incentives to a question regarding voting turnout. Figure B.7 — taken from the published version of the paper — shows the details of the randomization design. We replicate Table 1 (Columns 1 and 3) in the original paper including the interaction terms across treatments. Since the third and fourth randomization only take place after the respondent opens the door (which is the outcome we focus on) we focus on the first three dimensions. However, the paper does not mention that the estimates in the main tables (e.g., Table 1) should be interpreted as weighted averages of causal effects with respect to different counterfactuals.

Figure B.7: Factorial design in DellaVigna et al. (2016)

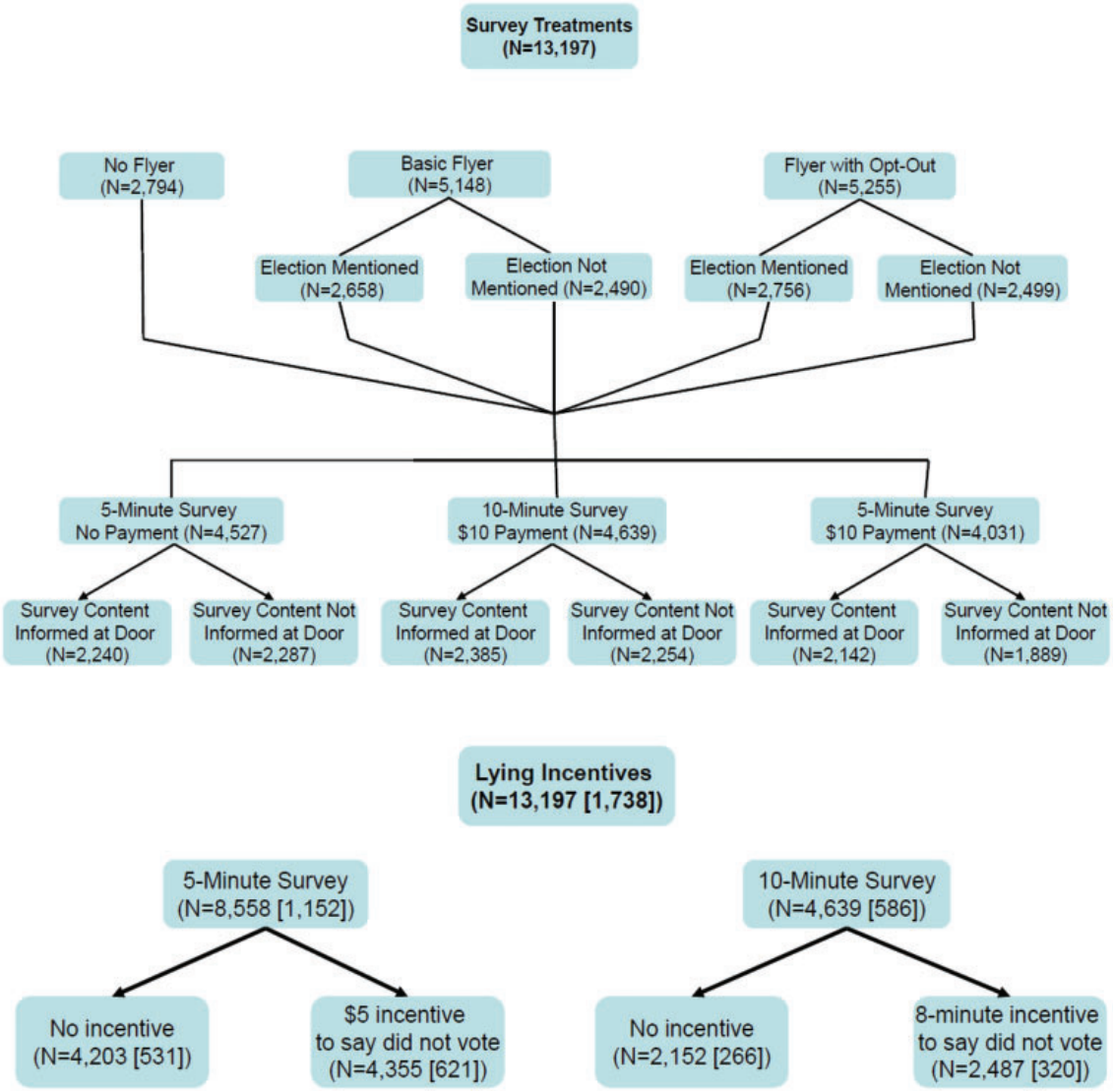


FIGURE 3

Experimental treatments

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

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

B.20 Contract Structure, Risk-Sharing, and Investment Choice

Fischer (2013) analyzes a field experiment in which individuals are assigned to a random group across two dimensions. In the first dimension, individuals are assigned to one of five contracts: autarky, individual liability, joint liability, joint liability with approval rights, and equity. In the second dimension, all of the financial contract

treatments except for autarky were also randomized across two monitoring regimes: perfect and imperfect public monitoring. The paper uses the long model throughout. We re-estimate Table VIII in the paper and record the effect of the treatments (and their interactions) on the total transfers (i.e., Column 1, 5, and 9).

B.21 Self-Control at Work

Kaur et al. (2015) analyze a field experiment in which data entry workers are assigned to different contract/payment structures across two dimensions. First, employees were randomized into three payday groups, which were paid in the evenings of Tuesday, Thursday, and Saturday, respectively, for work completed over the previous 7 days. The second dimension changed the contract structure across six different options. The main tables in the paper (e.g., Tables 2 and Table 4) estimate the short model. The paper does not mention that the estimates based on the short model must be interpreted as weighted averages of treatment effects with respect to different counterfactuals. While some of the tables look at some of the interaction effects (e.g., Table 7), they group treatments together when they do this. We re-estimate the treatment effects on productivity, attendance, and earnings.

B.22 Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial

Cohen et al. (2015) analyze a field experiment with three treatment arms are: (i) ACT subsidy at 3 levels, (ii) RDT subsidy, and (iii) whether RDT is provided free of cost at the time of purchase. The paper estimates the long model throughout. We re-estimate Table 2 using the long model.

B.23 Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia

Blattman et al. (2017) analyze a field experiment with a 2×2 design. Along one dimension participants were randomly assigned to an offer of cognitive-behavioral therapy. Along the second dimension, participants were randomly assigned $200 grants. The main tables in the paper estimate the long model. We re-estimate Table 2 using the long model.

B.24 Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors

Khan et al. (2015) analyze an experiment in which tax collectors are paid for performance. This experiment features a 4×2 design. In the first dimension, units are assigned to either control, information only, or three different bonus schemes (+ information). In the second dimension, units are assigned to either control or performance pay for senior tax officials. The results for the second randomization (i.e., performance pay for senior officials) are not in the paper. In addition, the interactions are not included in the estimating equations. The data are not available in the journal's website, but are available on the author's website.[5]

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

B.25 What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods

Balafoutas et al. (2013) analyze a field experiment about taxi rides in Athens, Greece. The experiment is set up to measure fraud and to examine the influence of passengers' observable characteristics on fraud. The experiment vary the characteristics of passengers different taxi drivers got along two dimensions. First, passengers appear to be either local, non-local natives, or foreigners. Passengers in the roles of locals and non-local natives spoke in Greek, whereas passengers in the role of foreigners spoke in English. Passengers in the role of non-local natives and foreigners asked the driver whether he knew the destination, adding as an explanation for asking that they were not familiar with the city. In addition, each passenger also appeared to be either high- or low-income. Passengers intended to be perceived as having high income were dressed in a suit and carried a briefcase, whereas low-income passengers were dressed casually and carried a backpack. Figure B.8 — taken from the published version of the paper — shows the details of the randomization design. The paper does not mention that the estimates in the main tables (e.g., Table 5) should be interpreted as weighted averages of treatment effects with respect to different coun-

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

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

terfactuals. None of the tables in the main paper or the appendix estimate the long model. We re-estimate Table 5 (Columns 1-3) in the original paper including the interaction terms across treatments.

Figure B.8: Factorial design in Balafoutas et al. (2013)

TABLE 1
Treatments and locations in the experiment

[A] Treatments and number of observations

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

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

B.26 How Do Voters Respond to Information? Evidence from a Randomized Campaign

Kendall et al. (2015) study a field experiment with a 3 $\times$ 2 design in which voters are given information in different ways. In the first dimension, potential voters are randomized across a "valence flyer", a "ideology flyer", or control. In the second dimension, if they received a flyer this is randomized by both direct mail and phone calls or by direct mail only. Explicitly, they "randomly divided the 95 precincts into four groups: (i) 24 precincts received the valence message; (ii) 24 precincts received the ideology message; (iii) 24 precincts received both messages; (iv) 23 precincts received no message (control group). Furthermore, we randomly split the first three groups into two subgroups: in the first, the treatment was administered by both direct mail and phone calls (12 precincts); in the second, by direct mail only (12 precincts)." The main tables in the paper estimate the long model. We re-estimate Table 3 using the long model.

B.27 Why the Referential Treatment? Evidence from Field Experiments on Referrals

Pallais & Sands (2016) analyze three field experiments in an online labor market to study why referred workers are more likely to be hired than non-referred workers.

The same sample is randomized in three dimensions (the three experiments). The paper does not mention that the estimates in the main tables should be interpreted as the weighted average of treatment effects with respect to different counterfactuals. None of the tables in the main paper or the appendix estimate the long model. The data are not available online.