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

B.1 Short description of each paper with a factorial designs

B.1.1 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.9 — 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.9: 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.1.2 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.10, 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.10: 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.1.3 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.11 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.[30].

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

Figure B.11: Factorial design in Duflo et al. (2011) and Duflo et al. (2015b)

Figure 1
Experimental Design: The Extra-Teacher Project

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

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

B.1.4 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.1.5 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.1.6 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.1.7 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.1.8 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.1.9 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.[31] Figure B.12 — 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

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

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.

Figure B.12: Factorial design in Haushofer & Shapiro (2016)

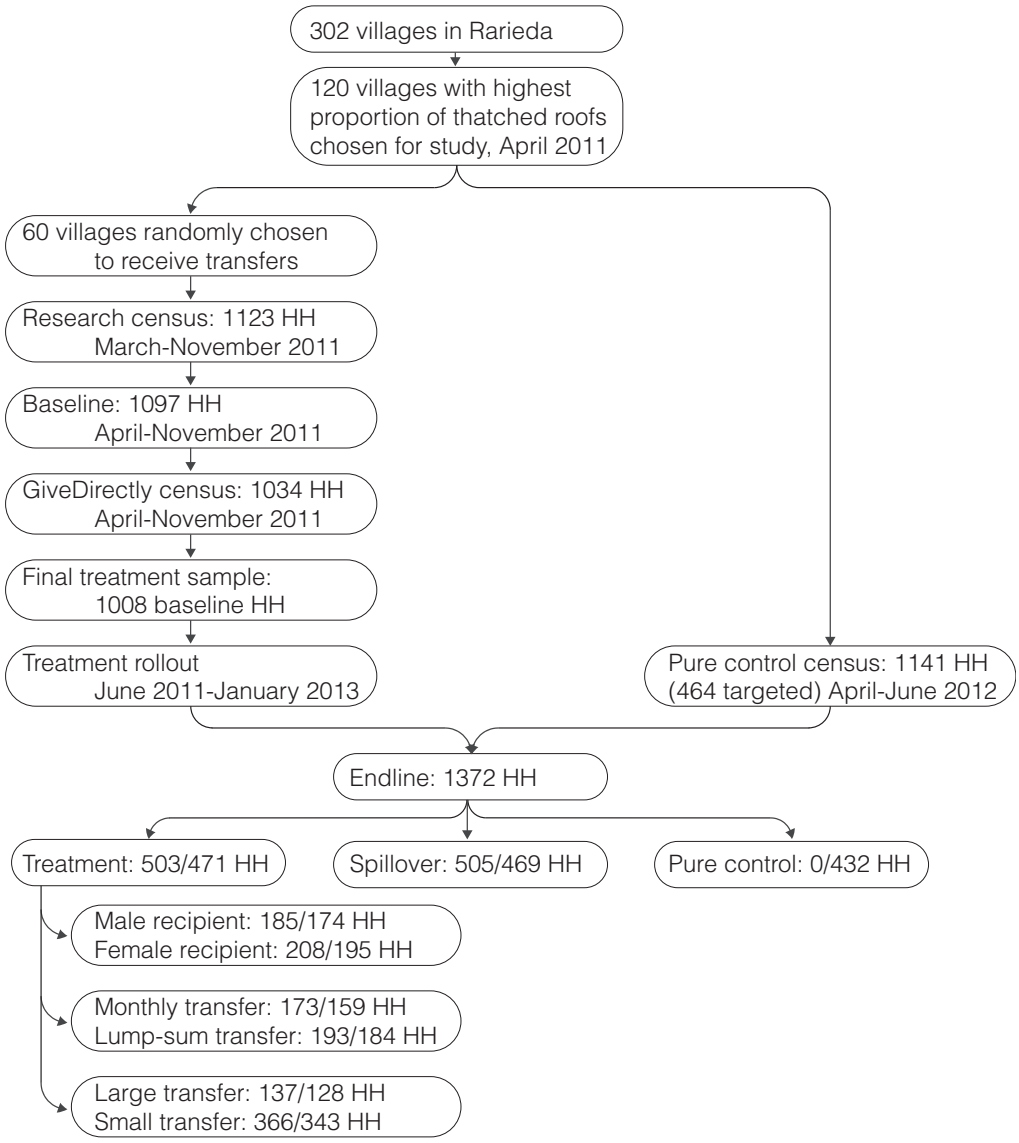

FIGURE I

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.1.10 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 the targeting is done using a proxy-means test, in some the targeting is done by the community, and in some is a hybrid of both. In "community" and "hybrid" villages the treatments had several variations: In some villages, the meetings took place during the day, in others at night. In some, the "elite" of the village took the decision, in some, it was the whole community. In some, the 10 poorest households were primed by the meeting facilitator, in some, there was no priming. Explicitly, the paper states "We designed several subtreatments in order to test three hypotheses about why the results from the community process might differ from those that resulted from the PMT treatment: elite capture, community effort, and within-community heterogeneity in preferences." Figure B.13 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.13: 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.1.11 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.[32] However, the paper does not mention

[32]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).

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 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.1.12 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.[33] 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.1.13 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.

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

B.1.14 Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies

Jakiela & Ozier (2015) analyze a experiment to measures 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.1.15 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., the ethnicity or gender of the applicant (cf. Riach and Rich 2002; Carlsson and Rooth 2007). However, in our experiment, we used a more general approach by randomly varying

several characteristics. This allows us to measure the labor market return of different skills and attributes (cf. Bertrand and Mullainathan 2004; Rooth 2011)." The paper does not mention that the 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.1.16 Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market

Allcott & Taubinsky (2015) reports on two experiments, both of which have a 2×2 designs. Figure B.14 — 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.14: Factorial design in Allcott & Taubinsky (2015)

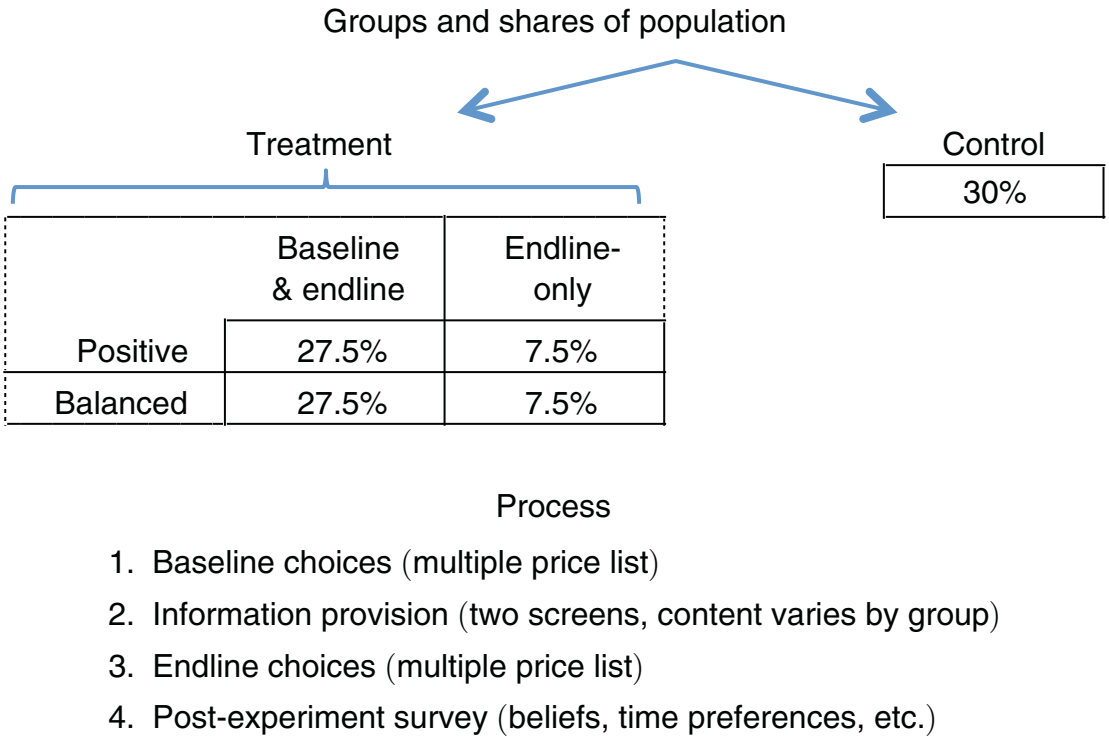

FIGURE 1. TESS EXPERIMENTAL DESIGN

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

Flory et al. (2014) analyses two experiments. The first experiment uses a $2\times6\times2$ 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.1.18 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.1.19 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.15 — 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.15: Factorial design in DellaVigna et al. (2016)

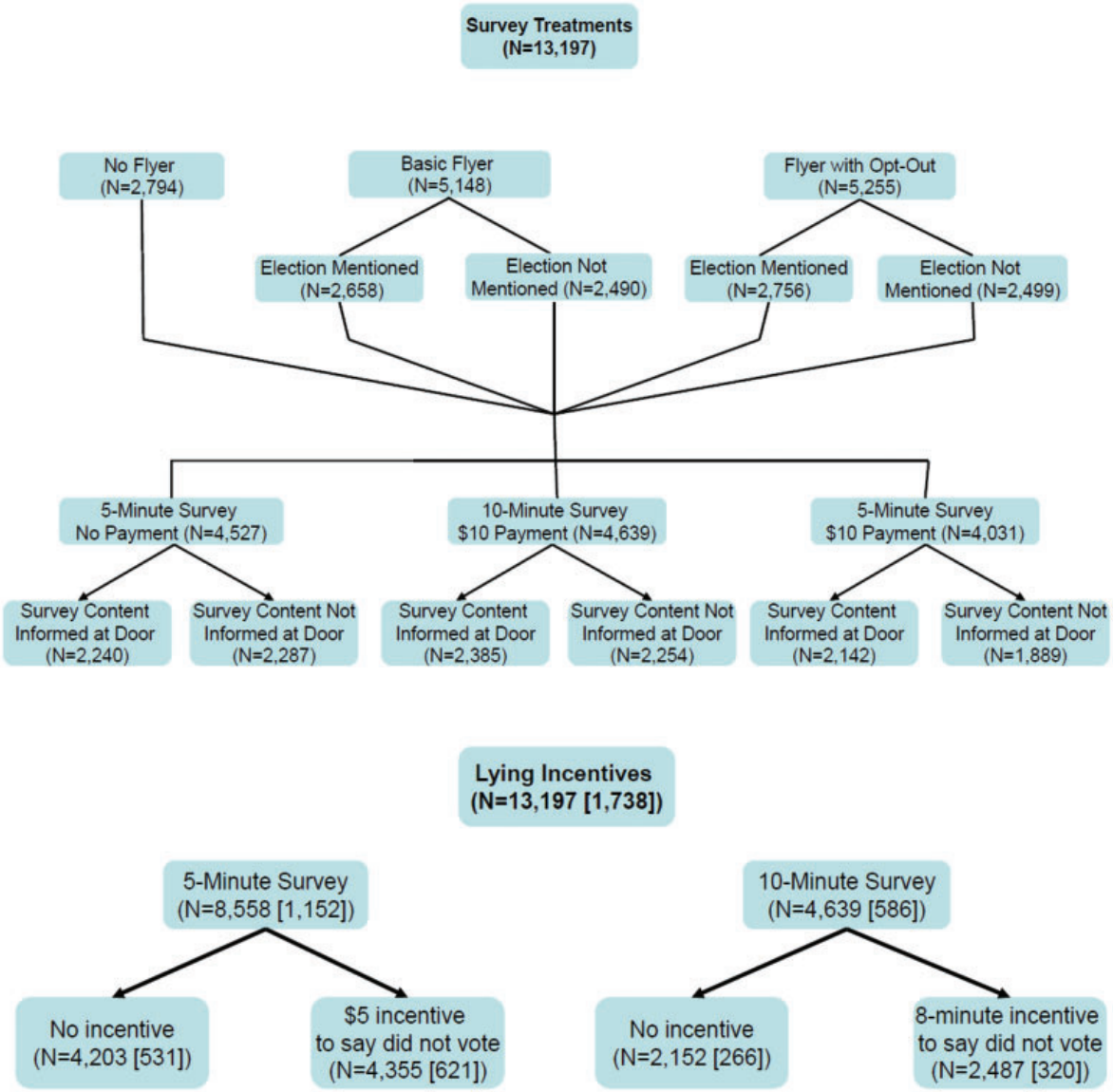

FIGURE 3

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.1.20 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.1.21 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.1.22 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 time of purchase. The paper estimates the long model throughout. We re-estimate Table 2 using the long model.

B.1.23 Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia

Blattman et al. (2017) analyzes a field experiment with a 2×2 design. Along one dimension participates 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.1.24 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.[34]

The second treatment (incentives for senior officials) only took place during the second year of the experiment. The paper does not mention that the treatment effects in the main tables (e.g., Table 3) should be interpreted as the weighted average over the "senior officials treatment status". None of the tables in the main paper or the appendix estimate the long model. Thus, we re-estimate all the results in Columns 4 to 6 of Table 3 (Panel B) including all the interactions between treatments.[35]

B.1.25 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.

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

[35] 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.

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.16 — 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 counterfactuals. 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.16: 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.1.26 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.1.27 Why the Referential Treatment? Evidence from Field Experiments on Referrals

Pallais & Sands (2016) analyzes three field experiments in an online labor market to study why referred workers are more likely 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.