Choosing What to Learn: Experimental Design when Combining Experimental with Observational Evidence*

Aristotelis Epanomeritakis† Davide Viviano‡

November 30, 2025

Abstract

Experiments deliver credible treatment-effect estimates but are typically localized to specific sites, populations, or mechanisms. A growing literature therefore uses these estimates as inputs for broader policy questions, combining them with reduced-form or structural observational evidence. We develop a unified framework for choosing which experiment to run and how to run it in this setting. We evaluate designs using a minimax proportional-regret criterion that compares any candidate design to an oracle that knows the observational study bias and jointly chooses the design and estimator. This yields a transparent bias–variance trade-off that does not require the researcher to specify a bias bound and relies only on information already needed for conventional power calculations. We illustrate the framework by (i) designing cash-transfer experiments aimed at general equilibrium schooling effects and (ii) optimizing site selection for microfinance interventions.

*We thank Dmitry Arkhangelesky, Raj Chetty, Kevin Chen, Larry Katz, Hiro Kaido, Gabriel Kreindler, Konrad Menzel, Jesse Shapiro, Elie Tamer, Jaume Vives-i-Bastida for helpful comments and discussion. We thank Giacomo Opocher and Nabin Poudel for exceptional research assistance. Davide Viviano acknowledges support by the Harvard Griffin Fund in Economics and NSF Grant SES 2447088. All mistakes are our own.

†Department of Economics, Harvard University. Email address: aristotle_epanomeritakis@fas.harvard.edu.

‡Department of Economics, Harvard University. Email address: dviviano@fas.harvard.edu.

1 Introduction

Randomized controlled trials (RCTs) have improved empirical economics by providing internally valid causal estimates. However, feasibility constraints often confine trials to localized effects such as the effect in a specific site or subpopulations, or of a particular mechanism. These effects, although useful, are often not sufficient to answer broader questions about external validity and general equilibrium (GE) effects of at-scale interventions. For instance, the median evaluation of experiments in top economics journals is only "representative of a population of 10,885 units, studied a treatment delivered to 5,340 units, and randomized in clusters of 26 units per cluster" (Muralidharan and Niehaus, 2017).

In response, a growing literature in development economics (e.g., Attanasio et al., 2012; Bassi et al., 2022; de Albuquerque et al., 2025; Meghir et al., 2022), education (e.g., Allende et al., 2019; Larroucau et al., 2024), labor economics (e.g., Chetty et al., 2016), and industrial organization (e.g., Allcott et al., 2025; Katz and Allcott, 2025) combine experiments with external evidence—observational estimates and/or results from experiments in other contexts—to estimate complex counterfactuals that no single experiment can identify. In our review of AEA journals over the past decade, more than 30% of experimental papers complement RCT evidence with observational inputs, either reduced-form or structural (Figure 12). This trend poses a design question: given constraints on feasible experiments, which experiments should be run (and how) to estimate such counterfactuals?

Our main contribution is a framework for designing experiments combined with external evidence. By external evidence we mean reduced-form or structural estimates based on observational data, and/or results from experiments conducted in other contexts. We provide a procedure that selects which experiments to run subject to a budget (e.g., which treatment arms and/or sub-populations to include), sets their precision via sample allocation, and prescribes how to combine experimental and external evidence. External input may be biased in ways not known ex ante (e.g., due to confounding or lack of external validity).

For an illustrative example, consider a government piloting a cash-transfer program in a small set of districts to increase children's school attendance. The trial delivers a local average effect, but policy decisions often require counterfactuals at scale, allowing prices and wages to adjust (e.g. Egger et al., 2022; Todd and Wolpin, 2006). Researchers therefore combine experimental evidence with a supply-demand model that leverages observational data to map local impacts into economy-wide outcomes. The design question is twofold: (i) which parameters are the most valuable to learn experimentally given feasibility constraints on the size and type of experiment and (ii) how to allocate sample across sites and arms so that, once combined with external inputs, we precisely estimate the effect at scale.

The first step is to specify the estimand and the underlying parameters. Define $\tau(\theta)$ the (policy-relevant) object of interest, where τ is a known smooth function and θ is a vector of unknown parameters; τ may be scalar or vector-valued. Researchers also observe external baseline estimates of θ—henceforth *observational estimates*—which may be biased. Each feasible experiment can identify a subset of parameters without bias. This parameterization makes explicit which components of θ are learned experimentally and which may rely on biased observational evidence; not all components need (or can) be learned experimentally.

Returning to the cash-transfer example, $\tau(\theta)$ is the effect on schooling when all poor households in rural Kenya receive the subsidy. The parameter vector θ bundles (i) the direct schooling response to a conditional cash transfer (CCT), (ii) the income effect of transfers in partial equilibrium, and (iii) wage adjustments that shift the returns to schooling. Due to cost constraints, researchers may only run small, partial-equilibrium experiments and then combine these with prior evidence to recover $\tau(\theta)$. They may choose between a CCT treatment arm to estimate the direct effect, or an unconditional-cash arm (UCT) to measure the income elasticity, and extrapolate the GE effects using evidence from a previous study in Mexico (Todd and Wolpin, 2006). (When site selection is also part of the design, θ additionally collects site-specific effects and $\tau(\theta)$ targets the average effect across sites.)

The second step is to design the experiment: (i) selecting which experiment to run; (ii) allocating sample sizes across arms; and (iii) choosing how to combine observational and experimental evidence. We allow flexibility in (iii): researchers may choose to either optimally combine both sources when available or rely solely on experimental estimates. Parameters not identified experimentally are based on observational estimates.

A natural starting point is to design the experiment to minimize the mean-squared error (MSE) for estimating $\tau(\theta)$, perhaps worst-case over the observational bias. In practice, however, the size and even the direction of bias are unknown (before the experiment). A worst-case approach would therefore be overly-conservative and disregard information about the variance. We instead adapt the definition of proportional (adaptation) regret previously studied for combining two estimators generated by a fixed design (Armstrong et al., 2024; Tsybakov, 1998) and generalize it to the experimental design problem (with multiple parameters). Here, the regret is the ratio of the worst-case researcher's mean-squared error relative to an oracle that knows the worst-case bound on the observational bias and chooses both the *design* and the estimator. Taking its supremum over the worst-case bias yields a robust procedure without requiring a prior about the bias (or on its bound).

Our main result gives an explicit novel characterization of proportional regret. For any design and shrinkage weights that combine the experimental and observational point estimates, the regret is the maximum of two normalized components. The first is a *variance*

regret: the estimator's sampling variance under the chosen design and weights, divided by the smallest attainable variance. The second is a *bias regret*: the worst-case squared bias induced by using external evidence, divided by the smallest feasible worst-case bias over the class of designs. The variance regret depends on the (expected) variance-covariance matrix of the observational and experimental estimates implied by the design, the shrinkage weights, and the sensitivity of the policy parameter, i.e., the gradient of τ with respect to the parameters. This gradient is evaluated, under mild conditions, at the observational estimates.[1] The bias regret depends only on the sensitivity vector and the shrinkage weights, known ex-ante.

This characterization makes the bias-variance trade-off transparent. At the optimum, the two normalized components tend to be equalized: when the bias component dominates, the design prioritizes high-sensitivity parameters; when variance dominates, the procedure invests sample where it most reduces variance and relies more on the most precise observational inputs. This yields a nested procedure that chooses jointly the estimator, the sample allocation and the experiment to run, trading off precision with sensitivity of the policy estimand. The resulting design and estimator only require knowledge of the expected variance-covariance matrix (standard in experimental design problems, Gerber and Green, 2012) and of the observational estimates. It can then be reported in a pre-analysis plan. That is, the inputs for the design are the same that would be required by standard power-analysis plans. Optimization can be solved using off-the-shelf routines for mixed-integer quadratic programs.

We then extend the model to general method-of-moments estimators that combine experimentally identified moments with observational moments. As in the baseline case, observational moments may incur unknown misspecification. We optimize (i) which moments to identify experimentally and (ii) how to combine experimental and observational moments to estimate the counterfactual of interest. Our main result continues to hold for estimators are (asymptotically) linear in the moments. Linearity occurs (under smoothness conditions) as long as the ratio of the bias to standard error is no faster than $n^{1/4}$, with n denoting the observational sample size; standard local-asymptotics with the bias and standard error of the same order is a special case. Our results extend to general moment selection problems.

From a theoretical perspective, our key insight is that, for any norm used to define the ambiguity set for bias and for estimators that are asymptotically linear in the bias, the worst-case mean-squared error decomposes into a variance term plus the squared dual norm of a sensitivity vector multiplied by the squared radius of the ambiguity set. This struc-

[1]For this result to hold, we assume that $\tau(\theta)$ is twice differentiable with bounded derivatives. For nonlinear $\tau(\theta)$ in θ, we also require that the observational estimates bias is local to zero in the spirit of Andrews et al. (2020); Armstrong and Kolesár (2021); Bonhomme and Weidner (2022), where its rate of convergence can be the same, faster, or even slower than the estimators' standard error. Section 4.3 provides details.

ture implies that proportional regret is quasi-convex in the radius and yields a transparent variance-bias trade-off. This characterization is specific to our ex-ante design problem and does not arise in problems with a fixed design where the goal is to design the optimal ex-post estimator as in Armstrong et al. (2024), since there the estimator depends nonlinearly on the bias radius and the resulting regret objective is not quasi-convex in the design.

Finally, we study incomplete models in which $\tau(\cdot)$ is only partially identified. We define regret with respect to the worst-case length of the estimated identified set, accounting for sampling uncertainty. Our variance–bias trade-off characterization continues to hold, with the bias now reflecting the sensitivity of lower and upper bound functions. For known $\tau(\cdot)$ where we are interested in minimizing confidence-interval length, MSE-based regret optimal design coincide with optimal designs for confidence interval length.

We illustrate the framework in the cash-transfer application, when the researcher uses evidence from Mexico's PROGRESA program (Attanasio et al., 2012; Todd and Wolpin, 2006). We are concerned that these preliminary estimates may lack external validity in Kenya. We compare three designs: (i) a CCT arm to estimate the direct schooling effect; (ii) an unconditional cash transfer (UCT) arm to estimate income effects; and (iii) a two-arm design that allocates sample across CCT and UCT under a common budget. Running a single arm captures settings with fixed costs per arm; allowing both arms with budgeted allocation captures cases with non-binding fixed costs but binding variable costs. When both arms are feasible, the regret-optimal allocation puts most participants in the CCT and a small fraction in the UCT, reflecting the target's greater sensitivity to the direct effect despite the CCT's higher variance. If only one arm can be run, the choice hinges on sample size: for very small n, the lower-variance of the UCT arm makes this preferred over a (too noisy) CCT arm; once $n \geq 500$, the more informative but noisier CCT yields lower regret. Relative to an oracle that knows the bias, the optimal design closely tracks the oracle.

As a second application, we study where to run a microfinance experiment in rural India, integrating evidence from earlier nonrandomized microfinance introductions. We start from the observational estimates in Banerjee et al. (2024) and design an experiment that selects one or more areas for randomization. Because Banerjee et al. (2024) also implement a separate randomized expansion, we can calibrate performance of our design (that is agnostic of the bias) by comparing each candidate design's MSE to that of an oracle that knows the bias, using the experimental estimates to calibrate the bias. Compared to a benchmark that randomly chooses areas and splits sample equally, we reduce MSE by roughly 20%.

Related literature This paper links the experimental-design literature—which mostly focuses on settings where all parameters are identified within the experiment and leaves aside questions of data combination—to recent work that integrates experimental and (reduced-

form or structural) observational evidence to extrapolate effects in complex scenarios.

Recent advances for experimental design include balancing and variance-minimizing allocations (Bai, 2019; Bertsimas et al., 2015; Cytrynbaum, 2021; Kallus, 2018; Tabord-Meehan, 2018), adaptive designs for policy choice (Cesa-Bianchi et al., 2025; Kasy and Sautmann, 2019; Russo et al., 2018), and experimental design under correct model specification (Chaloner and Verdinelli, 1995; Chaudhuri and Mykland, 1993; Higbee, 2024; Kasy, 2016; Kiefer and Wolfowitz, 1959; Reeves et al., 2024; Silvey, 2013; Viviano, 2020). Traditional work on experimental design for robust-model estimation either focused on testing competing models (Atkinson and Fedorov, 1975; López-Fidalgo et al., 2007), or on using a-priori knowledge of (worst-case) bias for the design of an experiment (Box and Draper, 1959; Sacks and Ylvisaker, 1984; Tsirpitzi et al., 2023; Wiens, 1998). All these references leave aside questions about observational data combination. In the context of minimizing the variance of an experiment, Rosenman and Owen (2021) proposes to use observational studies to construct high-confidence bounds on the experimental variance, and then focus on variance-optimal stratification. This problem differs from our design problem and analysis, which instead studies the use of observational studies in combination with experiments for extrapolation.

In summary, we complement this literature by studying which experiment to run (and how) when not all parameters relevant for the target estimand can be learned experimentally.

Our minimax regret criterion connects to a long-standing decision-theoretic tradition for experimental design. References include Manski and Tetenov (2016), Banerjee et al. (2020), Manski (2004), Dominitz and F. Manski (2017), Olea et al. (2024), Hu et al. (2024), Breza et al. (2025) among others. These references focus on settings where researchers have only access to experimental variation, instead of combining it with observational evidence, motivating different designs and objective functions. In the context of site-selection, we complement literature that leverages correctly-specified models (Abadie and Zhao, 2021; Gechter et al., 2024) by allowing for misspecification in observational estimates.

A related line of work analyzes estimation under misspecification (Andrews et al., 2025, 2020; Armstrong et al., 2024; Armstrong and Kolesár, 2018; Bonhomme and Weidner, 2022; Christensen and Connault, 2023; James et al., 1961), and combining existing experiments with observational studies (Athey et al., 2020, 2025; Bhattacharya, 2013; de Chaisemartin and D'Haultfœuille, 2020; Dutz et al., 2021; Gechter, 2022; Kallus et al., 2018; Rambachan et al., 2024; Rosenman et al., 2022). Our definition of sensitivity of the estimand to each parameter builds on sensitivity analysis in Andrews et al. (2020). Unlike both strands of this literature where the design is fixed and researchers optimize over the choice of the estimator only, here we optimize the design itself. This changes the definition of regret (and optimization), computed against the best design-estimator combination.

2 Baseline scenario: problem description

Consider a researcher interested in an arbitrary target estimand $\tau(\theta) \in \mathbb{R}$, indexed by a low-dimensional unknown parameter vector $\theta \in \mathbb{R}^p$ and a known mapping τ.

The goal is to construct an estimator $\hat{\tau}$ that accurately approximates $\tau(\theta)$ by combining existing evidence (e.g., observational studies) with experimental variation designed by the researcher. Our question is how to design such experiments under feasibility constraints.

Basic setup We assume that researchers have access to estimators (and their variance) of θ denoted as $\tilde{\theta}^{\text{obs}} \in \mathbb{R}^p$. We impose no restrictions on how $\tilde{\theta}^{\text{obs}}$ is formed: it can based on arbitrary exclusion restrictions implied by an economic or statistical model. However, $\tilde{\theta}^{\text{obs}}$, defined as *observational estimates*, have unknown biases collected in a vector $b \in \mathbb{R}^p$. Examples include an estimate from an instrumental variable regression that may fail the exclusion restriction, or an estimate from a country different from the one of interest that may lack external validity. Researchers can also collect experimental evidence for a *subset* of parameters, producing estimates whose bias equals zero by design.

Setting 1. For a subset $\mathcal{E} \subseteq \{1, \ldots, p\}$ and a known positive-definite matrix $\Sigma(\mathcal{E}) \in \mathbb{R}^{(p+|\mathcal{E}|)\times(p+|\mathcal{E}|)}$, define $\tilde{\theta}^{\text{obs}} \in \mathbb{R}^p$ an observational estimate and $\tilde{\theta}^{\text{exp}}_{\mathcal{E},\Sigma} \in \mathbb{R}^{|\mathcal{E}|}$ an experimental estimate each satisfying

$$\mathbb{E}\left[\tilde{\theta}^{\text{obs}}\right] - \theta = b, \qquad \mathbb{E}\left[\tilde{\theta}^{\text{exp}}_{\mathcal{E},\Sigma}\right] - \theta_{\mathcal{E}} = 0,$$

for an unknown bias vector $b \in \mathbb{R}^p$. Moreover, define its joint variance-covariance matrix as $\mathbb{V}\begin{pmatrix} \tilde{\theta}^{\text{obs}} - \theta \\ \tilde{\theta}^{\text{exp}}_{\mathcal{E},\Sigma} - \theta_{\mathcal{E}} \end{pmatrix} = \Sigma(\mathcal{E})$. Given $(\mathcal{E}, \Sigma)$, consider a class of linear plug-in estimators

$$\hat{\tau}_\gamma \equiv \tau(\hat{\theta}(\gamma)), \qquad \hat{\theta}_j(\gamma) = \begin{cases} \gamma_j\, \tilde{\theta}^{\text{exp}}_j + (1-\gamma_j)\, \tilde{\theta}^{\text{obs}}_j, & j \in \mathcal{E}, \\ \tilde{\theta}^{\text{obs}}_j, & j \notin \mathcal{E}, \end{cases}$$

where $\gamma = (\gamma_j)_{j\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ are shrinkage weights that can be an (implicit) function of $(\mathcal{E}, \Sigma)$.

Here, $(\mathcal{E}, \Sigma(\mathcal{E}))$ characterizes the experimental design (which parameters are learned and with what precision). When clear, we omit the subscript Σ on $\tilde{\theta}^{\text{exp}}$. We assume $\Sigma(\mathcal{E})$ is known, which is standard in experimental planning (Gerber and Green, 2012) (in practice, any consistent estimator, e.g., from pilot studies would suffice).

The class of experiments and estimators is assumed to be in a user-specific set $\mathcal{D}$.

Assumption 1 (Feasible experiments and estimators)**.** We write compactly $(\mathcal{E}, \Sigma, \gamma) \in \mathcal{D}$ indicating $\mathcal{D}$ the set of feasible design (i.e., choice of $\mathcal{E}$ and $\Sigma(\mathcal{E})$) and estimators (i.e., γ). The set $\mathcal{D}$ can be arbitrary as long as (i) each feasible Σ has uniformly bounded entries and is strictly positive definite; (ii) $\gamma_j = 1$ for all $j \in \mathcal{E}$ is one admissible choice (among potentially other values).

We let $\mathcal{E}, \Sigma(\mathcal{E})$ be within arbitrary constraint sets that may encode feasibility or budget constraints. In our framework, restrictions on the experiments researchers can run $\mathcal{E}$ and on their precision $\Sigma(\mathcal{E})$ can take any desired form. In most applications, we think of such constraints as arising from fixed costs of running an experiment and/or constraints on their power (Athey and Imbens, 2017; Duflo et al., 2007; List et al., 2011). We assume that $\Sigma(\mathcal{E})$ is uniformly bounded, which implies that once we commit to learn a set of experimental parameters $\tilde{\theta}^{\text{exp}}_{\mathcal{E}}$, their variance is finite, and strictly positive definite, therefore assuming that the variance is bounded away from zero.[2] The shrinkage estimator can be unrestricted (i.e., γ can take any value on the real line) or restricted (when for example researchers aim to only use experimental evidence when available, setting $\gamma_j = 1$ for $j \in \mathcal{E}$).

We extend our framework to more general methods of moments and multivariate target parameters $\tau(\theta) \in \mathbb{R}^q, q > 1$ in Section 4.1.

Remark 1 (Structure of Σ)**.** Because observational and experimental moments are typically computed from disjoint samples, in practice, it is natural to model the covariance as block diagonal, $\Sigma = \text{diag}(\Sigma_{\text{obs}}, \Sigma_{\text{exp}})$. The first block corresponds to observational estimates (fixed and not subject to design choices), while the second corresponds to experimental estimates, and it is a function of the sample size allocated to treatment arms. This is a special case of our framework which can be encoded in the constraint set $\mathcal{D}$. When researchers have a prior about the experimental variance-covariance matrix, we interpret Σ_{exp} as its expectation.[3] □

Research questions Our design problem can be described in three steps:

1. *Preliminary step: weights.* For each feasible $\mathcal{E}, \Sigma(\mathcal{E})$, choose γ to combine experimental and observational estimates on the selected coordinates. (If researchers restrict their estimators to only use experimental evidence when available, set $\gamma = 1$.)

[2]For this latter condition to hold in an asymptotic framework with growing sample size, θ and $\tilde{\theta}$ can be defined without loss as parameters of interest after appropriately rescaling by the square-root of the sample size; in this case Σ denotes the asymptotic variance. See Section 4.3 for details.

[3]In this case, we interpret the MSE in Equation (2) as the expected MSE under a common expectation over the experimental variance-covariance matrix, assuming that its expectation Σ_{exp} is not data-dependent. It is also possible to extend our framework by allowing Σ_{exp} to be a function of $\tilde{\theta}^{\text{obs}}$ provided that $\Sigma_{\text{exp}}(\tilde{\theta}^{\text{obs}}) \to_p \Sigma_{\text{exp}}(\text{plim}(\tilde{\theta}^{\text{obs}}))$, with $\Sigma_{\text{exp}}(\text{plim}(\tilde{\theta}^{\text{obs}}))$ denoting the common expectation of the variance-covariance experimental matrix. In this case our analysis should be interpreted in an asymptotic regime similar to what discussed in Section 4.3, which we omit for brevity.

2. *Middle step: precision.* For each feasible $\mathcal{E}$, choose $\Sigma(\mathcal{E})$ (i.e., sample allocation).

3. *Outer step: experiment choice.* Choose a feasible set of experiments $\mathcal{E}$.

Choosing γ (step 1) for a fixed design follows a long-standing tradition in econometrics and statistics (Andrews and Shapiro, 2021; Armstrong et al., 2024; Athey et al., 2025; Donoho, 1994). We study this question here in combination with the experimental design problem, step 2 and 3; this, as we show, will change the underlying optimization problem (also for γ). The central challenge is that performance depends on the unknown bias b.

Mapping τ A key feature of our framework is that researchers explicitly parameterize $\tau(\cdot)$, thereby pre-specifying and making transparent which biases drive the design problem; we see this as an advantage because it clarifies the role of the experiment in complementing existing evidence. The sensitivity of τ to θ, captured by its gradient $\omega(\theta) = \partial\tau(\theta)/\partial\theta$, determines how (to first order) the bias propagates onto the final estimand. For simplicity, we impose exact linearity, $\tau(\theta) = \omega^\top\theta$; all results extend to smooth, nonlinear τ via a first-order expansion.

Assumption 2 (First-order estimation error)**.** For any $(\mathcal{E}, \Sigma, \gamma) \in \mathcal{D}$, assume

$$\tau(\theta) - \tau\big(\hat{\theta}(\gamma)\big) = \sum_{j=1}^{p} \omega_j \left(\theta_j - \hat{\theta}_j(\gamma)\right), \tag{1}$$

for known weights $\omega \in \mathbb{R}^p$, with $|\omega_j| \in (0, \infty)$ for all j.

Assumption 2 holds exactly when $\tau(\cdot)$ is linear and serves as a first-order approximation for smooth non-linear $\tau(\cdot)$. Section 4.3 (and Example 3) extend the framework to non-linear estimands encompassing as special cases the local-misspecification setups in Andrews et al. (2020); Armstrong and Kolesár (2021). For non-linear τ, we replace ω with the gradient of τ evaluated at a preliminary observational estimate, i.e., $\omega = \frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\tilde{\theta}^{\text{obs}}} + o_p(1)$, which is typically available before the experiment is conducted. That is, building on references above, ω captures the first-order effect of the observational bias b on the bias of $\hat{\tau}$.[4]

Remark 2 (Experimental design with partially identified τ)**.** Although in our applications $\tau(\theta)$ is known and pre-specified, Section 4.2 studies settings where researchers only know lower and upper envolopes functions $\bar{\tau}, \underline{\tau}$, with $\tau(\theta) \in [\underline{\tau}(\theta), \bar{\tau}(\theta)]$ partially identified and focuses on minimizing the length of the confidence (partially identified) set. □

[4]To see how this assumption holds under a first-order approximation, define θ the parameter rescaled by $\sqrt{n}$ (and similarly its bias b_n, indexed by n), where $n^{-1/2}$ is the rate of convergence of the observational estimators. Then second-order effects are negligible (and Assumption 2 holds asymptotically) provided that $b_n^2/\sqrt{n} \to 0$. This implies that Assumption 2 holds when the asymptotic bias b_n is proportional to, smaller or even much larger than the asymptotic variance, provided that b_n does not grow at rate faster than $n^{1/4}$.

We conclude with three examples in simple two-parameter models.

Example 1 (Choosing the site for an experiment for external validity)**.** Gechter et al. (2024) study where to run an experiment. For illustration, consider two sites $j \in \{1, 2\}$ with site-specific ATEs θ_j and target the cross-site average $\tau(\theta) = \omega_1\theta_1 + \omega_2\theta_2$, with ω denoting the population density in each site. Let $\tilde{\theta}^{\text{obs}} = (\tilde{\theta}_1^{\text{obs}}, \tilde{\theta}_2^{\text{obs}})^\top$ denote observational estimates obtained in Gechter et al. (2024) from a structural model and potentially biased due to misspecification: $\mathbb{E}[\tilde{\theta}^{\text{obs}}] - \theta = (b_1, b_2)^\top$. A budget constraint allows an experiment in only one site, so that $\mathcal{E} \in \{\{1\}, \{2\}\}$. If site j is chosen, we obtain an unbiased $\tilde{\theta}_j^{\text{exp}}$; the other site $k \neq j$ remains observational. Given $\mathcal{E}$,

$$\widehat{\tau}_\gamma = \tfrac{1}{2}\Big[\gamma_j\, \tilde{\theta}_j^{\text{exp}} + (1 - \gamma_j)\, \tilde{\theta}_j^{\text{obs}} + \tilde{\theta}_k^{\text{obs}}\Big], \qquad k \neq j.$$

Our question is how to choose the set $\{j\}$ where to conduct the experiment (jointly with γ).

Example 2 (Choosing which survey to conduct)**.** Egger et al. (2022) study the efficacy of cash-transfer programs on the marginal propensity to consume (MPC). Measuring the MPC requires capturing both short- and long-run effects. Because survey rounds are limited, the authors complement experimental data that lacks short-run effects with auxiliary information from prior studies that use short-run surveys (collected in other regions; see Egger et al. (2022)). This raises the question of which survey to conduct (and with which frequency).

In stylized form, suppose researchers can observe, for $s \in \{1, 2\}$, potential outcomes $Y_s(t)$ denoting consumption in period s when measured t periods after the intervention. The authors have auxiliary estimates from previous studies,

$$\alpha \;=\; \mathbb{E}\big[Y_1(t = 1) - Y_1(t = \infty)\big], \qquad \beta \;=\; \mathbb{E}\big[Y_2(t = 1) - Y_2(t = \infty)\big],$$

and wish to estimate the total effect $\tau = \alpha + \beta$. Researchers consider two survey designs:

(i) *Early survey design* to estimate α precisely;

(ii) *Later survey* to estimate β precisely.

Our goal is to study which survey design to implement. (More complex designs with additional time periods or mixed precision across rounds are also possible here.)

Example 3 (Supply or demand experiment with non linear target)**.** Bergquist and Dinerstein (2020) conduct demand and supply experiments in food markets in Kenya. Suppose here we are interested in similar applications in Uganda. For exposition, consider a basic

linear demand and supply

$$Q^D = a - \beta_D P + u_D, \qquad Q^S = c + \beta_S P + u_S, \qquad \theta = (\beta_D, \beta_S)^\top,$$

with $\beta_D \neq -\beta_S$. Several estimands may be of interest; one of such estimands is the effect of a tariff t on prices

$$\tau \equiv \frac{\beta_S}{\beta_D + \beta_S}\, t.$$

Let $(\tilde{\beta}_D^{\text{obs}}, \tilde{\beta}_S^{\text{obs}})^\top$ denote baseline estimates from Kenya, that *may* lack external validity.

Due to cost constraints, $\mathcal{E} \in \{\{D\}, \{S\}\}$, i.e., we can learn either demand (estimating $\tilde{\beta}_D^{\text{exp}}$, with randomized price discounts) or supply (estimating $\tilde{\beta}_S^{\text{exp}}$, by introducing regulatory cost shocks on firms). Let

$$\big(\tilde{\theta}^{\text{obs}} - \theta\big) \;\sim\; \mathcal{N}\big(b, \Sigma^{\text{obs}}\big), \quad \tilde{\theta}^{\text{obs}} \equiv \sqrt{n}(\tilde{\beta}_D^{\text{obs}}, \tilde{\beta}_S^{\text{obs}})^\top, \quad \theta \equiv \sqrt{n}(\beta_D, \beta_S)^\top$$

with θ and $\tilde{\theta}$ denoting the parameter and estimator here rescaled by the square-root of the sample size, and b capturing bias. Our goal is to choose whether to conduct a supply or demand experiment in Uganda to estimate τ. For given estimators $\hat{\theta}$ that combine the estimates from Kenya with the chosen experimental estimate from Uganda under local asymptotics described in Section 4.3

$$\tau - \widehat{\tau} = \underbrace{\frac{1}{\sqrt{n}} \breve{\omega}(\text{plim}(\tilde{\beta}^{\text{obs}}))^\top (\theta - \widehat{\theta})}_{\text{main first order effect}} + \underbrace{o_p\left(\frac{1}{n^{1/2}}\right)}_{\text{small higher order effects}}, \qquad \breve{\omega}(\beta) = \frac{t}{(\beta_D + \beta_S)^2} \begin{pmatrix} -\beta_S \\ \beta_D \end{pmatrix},$$

where we can replace $\breve{\omega}$ with its consistent counterpart $\breve{\omega}(\tilde{\beta}^{\text{obs}})$. Our goal is to choose between a supply or demand experiment accounting for first-order bias (and variance) effect.

3 Robust experimental design

Next, we introduce an experimental design focusing on the mean-squared error (MSE) of the estimator $\hat{\tau}$, defined, for given bias level b and design $(\mathcal{E}, \Sigma(\mathcal{E}))$ as

$$\text{MSE}_b(\mathcal{E}, \Sigma, \gamma) \;=\; \mathbb{E}_{\mathcal{E},\Sigma,b}\big[(\widehat{\tau}_\gamma - \tau)^2\big], \tag{2}$$

where $\mathbb{E}_{\mathcal{E},\Sigma,b}$ denotes expectation under the data-generating process implied by $(\mathcal{E}, \Sigma(\mathcal{E}))$ and observational bias b.

Ideally, one would minimize MSE_b. However, because b is unknown, we consider an

uncertainty set given by an ℓ_∞-ball,

$$\mathcal{B}(B) \equiv \{b : \ \|b\|_\infty \leq B\}, \tag{3}$$

with $B \geq 0$ an upper bound on the largest coordinate-wise bias. In our framework, a key challenge is that biases may arise from multiple parameters. Here, for exposition we first focus on the ℓ_∞-ball given its interpretability. This norm imposes symmetric restrictions on the biases and it does not force a trade-off across coordinates (a large bias in one component need not be "offset" by a small bias elsewhere), which is attractive when biases may be positively correlated across parameters. The drawback is that worst-case solutions depend on the radius B which may be unknown in practice.

Section 4.1 generalizes the framework to arbitrary norms and Section 4.2 shows that our results for MSE also extend when minimizing confidence intervals' length.

If an oracle knew B, a natural choice would be to pick the minimax design

$$\text{MSE}^*(B) \ \equiv \ \inf_{(\mathcal{E},\Sigma,\gamma)\in\mathcal{D}} \ \sup_{b\in\mathcal{B}(B)} \ \text{MSE}_b(\mathcal{E},\Sigma,\gamma).$$

However, having to specify B can pose a large burden on the researchers and make the choice of the experiment sensitive to B. Therefore, we seek designs that perform as close as possible to the oracle that knows B, uniformly over the values of B, following a long-standing tradition in decision theory (e.g. Kitagawa and Tetenov, 2018; Manski, 2004; Manski and Tetenov, 2007; Montiel Olea et al., 2023). Specifically, we minimize the worst-case proportional regret

$$\mathcal{R}(\mathcal{E},\Sigma,\gamma) \ \equiv \ \sup_{B\geq 0} \ \frac{\sup_{b\in\mathcal{B}(B)} \text{MSE}_b(\mathcal{E},\Sigma,\gamma)}{\text{MSE}^*(B)},$$

defined *adaptation regret* by Armstrong et al. (2024) (building in turn on Tsybakov, 1998).

There are two key distinctions relative to such references using the adaptation regret. First (and most importantly) we optimize over the design itself rather than fixing it ex ante; regret is defined relative to an oracle that chooses ex-ante both the estimator and the design. Second, we allow biases to arise across multiple parameters (rather than a single one).

3.1 Optimal design

A natural question is whether the adaptation regret provides natural trade-offs between the variance and the bias. To do so, denote $\mathbb{V}_{\mathcal{E},\Sigma}(\cdot)$ the variance operator for given $(\mathcal{E},\Sigma)$.

Definition 1 (Variance regret)**.** Let

$$\alpha(\mathcal{E}, \Sigma, \gamma) \equiv \mathbb{V}_{\mathcal{E},\Sigma}(\hat{\tau}_\gamma), \qquad \alpha^\star \equiv \min_{\mathcal{E}\in\mathcal{S}} \min_{\Sigma(\mathcal{E})\in\mathcal{G}(\mathcal{E})} \min_{\gamma\in\mathbb{R}^{|\mathcal{E}|}} \alpha(\mathcal{E}, \Sigma, \gamma), \tag{4}$$

with $\alpha^\star$ denoting the smallest feasible variance. We will refer to $\alpha/\alpha^\star$ as the *variance regret*.

The variance regret denotes the ratio between the variance of a given design and estimator relative to the smallest achievable variance. We provide a similar definition for the bias. Let $|1-\gamma|$ define the absolute value of each entry of the vector $1-\gamma$.

Definition 2 (Bias regret)**.** Define

$$\beta(\mathcal{E}, \gamma) \equiv \Big(\|\omega\|_1 - \|\omega_{\mathcal{E}}\|_1 + |1-\gamma|^\top |\omega_{\mathcal{E}}| \Big)^2, \quad \beta^\star \equiv \min_{\mathcal{E}\in\mathcal{S}} \Big(||\omega||_1 - \|\omega_{\mathcal{E}}\|_1 \Big)^2,$$

with $\beta/\beta^\star$ defined as the *bias regret*.

To gain insight, note that $B^2\beta^\star = B^2\Big(||\omega||_1 - ||\omega_{\mathcal{E}}||_1\Big)$ denotes the smallest worst-case bias researchers can achieve for given choice of the experiment $\mathcal{E}$. Similarly, β denotes the worst-case bias for a given design and shrinkage estimator.

An ideal design, would set the variance equal to $\alpha^\star$ and bias equal to $\beta^\star$. Unfortunately, this may be infeasible as it might require different choices of experiments and estimators to achieve one or the other. This raises the question of how to trade-off these two components.

This trade-off is formally characterized in our main theorem below. We will use the convention throughout that $0/0 = 1$.

Theorem 1. *Consider Setting 1 and let Assumptions 1, 2 hold. Then, for any* $(\mathcal{E}, \Sigma, \gamma) \in \mathcal{D}$*,*

$$\mathcal{R}(\mathcal{E}, \Sigma, \gamma) = \max\left\{ \frac{\alpha(\mathcal{E}, \Sigma, \gamma)}{\alpha^\star}, \quad \frac{\beta(\mathcal{E}, \gamma)}{\beta^\star} \right\}.$$

Proof. See Appendix A.1. □

Theorem 1 provides a simple expression of the regret which does not require researchers to specify B. The key insight is that the adaptation regret in our context, leads to a (strictly) quasi-convex objective function, whose worst-case solution is the maximum between the variance and the bias regret. This provides a simple and intuitive trade-off between the two. This characterization differs from previous results on adaptation regret in Armstrong et al. (2024): there, the regret objective is not quasi-convex in B and therefore takes a different form, because the problem is ex-post shrinkage for a fixed design rather than the ex-ante experimental design problem studied here.

Specifically, provided that the class of designs $\mathcal{D}$ is sufficiently flexible (outside boundary solutions), it follows that at the optimum we equalize the variance and bias regret as in Figure 1

$$\frac{\alpha}{\alpha^\star} = \frac{\beta}{\beta^\star}.$$

In practice, this amount to allocate sample size to noisier treatment arms when the variance dominates, and select treatment arms with largest sensitivity when the bias dominates.

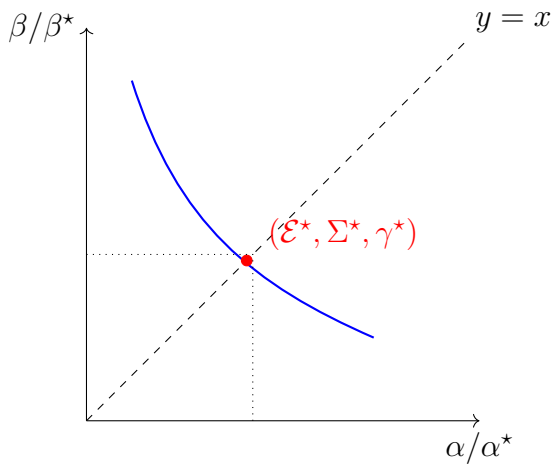

Figure 1: Each feasible design $(\mathcal{E}, \Sigma, \gamma)$ maps to a point $\big(\alpha/\alpha^\star,\ \beta/\beta^\star\big)$: the x–axis is the variance ratio and the y–axis is the worst–case bias ratio. The blue curve depicts the attainable frontier as we vary shrinkage γ and precision Σ. Level sets of the objective $\mathcal{R} = \max\{\alpha/\alpha^\star,\ \beta/\beta^\star\}$ are axis–aligned squares (the dotted inverted "L" shows the smallest such square touching the frontier). The minimizer outside boundary solutions is where the frontier meets the 45° line (red dot).

Theorem 1 provides us an immediate solution to the optimal design problem. We compute the estimator, the optimal variance and the experimental choice via backward induction. We first choose $\gamma^\star$ for each design $\mathcal{E}, \Sigma(\mathcal{E})$. We then choose Σ, whose choice depends on β through the shrinkage weights. We finally choose the set of experiments $\mathcal{E}$.

3.2 Optimization

Next, we provide an explicit optimization routine that can be solved using off-the-shelf software. We consider observational estimates independent of the experimental ones (but possibly dependent with each other), and experimental estimates independent of each other. This occurs when conducting experiments on independent samples different from the observational sample as in our applications. More general dependences are possible but omitted.

Algorithm 1 Regret-optimal experimental design

1: **Inputs:** target functional weights ω, variance parameters (v_j^2), observational covariance Σ_{obs}, per-unit costs (c_j), total budget n, and feasible experiment set $\mathcal{X}$, as in Section 3.2.
2: **Define variance, bias, and sample sizes.** For any (x, γ) with $x \in \mathcal{X}$ and $0 \le \gamma_j \le 1$, set $s_j = x_j\gamma_j$ and compute:
   - the sample allocation (6), and the definition of $\alpha(s)$ in (7);
   - the bias contribution $\beta(s)$ as defined below (7).
3: **Compute oracle quantities.** Obtain $\alpha^\star$ and $\beta^\star$ by solving the two optimization problems in the paragraph "Oracle solutions," i.e.
   - minimize $\alpha(s)$ over (x, γ, s) subject to $x \in \mathcal{X}$, $x_j \in \{0, 1\}$, $s_j = x_j\gamma_j$;
   - minimize $\sum_j x_j|\omega_j|$ over $x \in \mathcal{X}$ with $x_j \in \{0, 1\}$.
4: **Solve the minimax MIQP.** Solve (8)–(12) with a off-the-shelf solver.
5: **Outputs.** Let $(x^\star, \gamma^\star, s^\star, t^\star)$ denote an optimizer of (8)–(12). Report:
   - the regret-optimal experiment set $\mathcal{E}^\star = \{j : x_j^\star = 1\}$ and shrinkage weights $\gamma^\star$;
   - the implied $s^\star$ and regret factor $t^\star = \max\{\alpha/\alpha^\star, \beta/\beta^\star\}$;
   - the optimal sample sizes $n_j^\star$ obtained by plugging $s^\star$ into (6).

Notation and decision variables For given $(\mathcal{E}, \Sigma, \gamma)$, denote by Σ_{obs} the submatrix of Σ corresponding to the variance–covariance matrix of the observational estimates (possibly non-diagonal), and let $\mathbb{V}(\tilde{\theta}_j^{\text{exp}}) = v_j^2/n_j$ be the variance of the experimental estimator for component j, where n_j denotes the sample size allocated to that experiment. Let $x_j = 1\{j \in \mathcal{E}\} \in \{0, 1\}$ indicate whether the experiment identifies component j, and let $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the constraint set of feasible experiments. We optimize over (x, γ, n_j) under a total budget constraint $\sum_{j\in\mathcal{E}} c_j n_j = n$, where $c_j > 0$ denotes a per-unit cost for experiment.

Variance, bias and optimal sample size Under independence of experimental and observational estimators, the variance of $\tau(\theta)$ can be written as

$$\sum_j \omega_j^2 s_j^2 \frac{v_j^2}{n_j} \;+\; \big(\omega \odot (1 - s)\big)^\top \Sigma_{\text{obs}} \big(\omega \odot (1 - s)\big), \qquad s_j \equiv x_j\gamma_j, \tag{5}$$

where $\odot$ denotes the elementwise product. The sample sizes n_j enter only through the first (experimental) component. By the standard Neyman allocation with variable costs (Gerber and Green, 2012), the minimizer over (n_j) for fixed s is given by

$$n_j^\star \;=\; n \frac{|\omega_j| v_j s_j/\sqrt{c_j}}{\sum_k |\omega_k| v_k s_k/\sqrt{c_k}}. \tag{6}$$

Substituting (6) into the experimental variance and with a slight abuse of notation, let

$$\alpha(s) \;\equiv\; \frac{1}{n}\Big(\sum_j s_j|\omega_j|v_j\sqrt{c_j}\Big)^2 + \big(\omega\odot(1-s)\big)^\top \Sigma_{\text{obs}}\big(\omega\odot(1-s)\big), \quad \beta(s) \;\equiv\; \Big(\sum_j (1-s_j)|\omega_j|\Big)^2, \tag{7}$$

the variance and bias contributions.

Oracle solutions The next step is to compute $\alpha^\star$ and $\beta^\star$. These solve

$$\begin{aligned}
\alpha^\star = \min_{x,\gamma,s} \alpha(s), \qquad & \text{s.t. } 0 \le \gamma_j \le 1,\;\; x_j \in \{0,1\},\;\; x \in \mathcal{X}, \\
& 0 \le s_j \le \gamma_j,\;\; s_j \le x_j,\;\; s_j \ge \gamma_j + x_j - 1, \quad \forall j, \\
\sqrt{\beta^\star} = \min_{x} \sum_j x_j|\omega_j|, \quad & \text{s.t. } x \in \mathcal{X},\;\; x_j \in \{0,1\}, \quad \forall j.
\end{aligned}$$

The constraints on s are standard inequalities enforcing $s_j = \gamma_j x_j$ when $x_j \in \{0,1\}$ and $0 \le \gamma_j \le 1$. Because $\alpha(s)$ is convex quadratic and $\mathcal{X}$ typically admits a linear (or quadratic) description, the first problem is a mixed-integer convex quadratic program, while the second is a mixed-integer linear program. Both can be solved using off-the-shelf solvers.

Regret-optimal solution We are now ready to solve for the regret-optimal design. By Theorem 1, we want to minimize $\max\{\alpha/\alpha^\star, \beta/\beta^\star\}$. The corresponding optimization program can be written as

$$\min_{x,\gamma,s,t} \quad t \tag{8}$$

$$\text{s.t.} \quad \alpha(s) \;\le\; t\,\alpha^\star, \tag{9}$$

$$\beta(s) \;\le\; t\,\beta^\star, \tag{10}$$

$$0 \le \gamma_j \le 1, \quad x_j \in \{0,1\}, \quad j = 1,\dots,p, \tag{11}$$

$$0 \le s_j \le \gamma_j, \quad s_j \le x_j, \quad s_j \ge \gamma_j + x_j - 1, \quad j = 1,\dots,p, \tag{12}$$

$$x \in \mathcal{X} \quad \text{(constraint on feasible experiments)}. \tag{13}$$

The constraints (12) again enforce $s_j = x_j\gamma_j$ for $x_j \in \{0,1\}$ and $0 \le \gamma_j \le 1$. The variable t satisfies $t \ge \alpha/\alpha^\star$ and $t \ge \beta/\beta^\star$, so at the optimum we have $t = \max\{\alpha/\alpha^\star, \beta/\beta^\star\}$ as desired. Because both $\alpha(s)$ and $\beta(s)$ are convex quadratic functions, the optimization problem (8)–(12) is a mixed-integer quadratic program (MIQP) with linear objective and convex quadratic constraints, and can be passed to standard MIQP solvers.

3.3 Intuition with two parameters

To build further intuition, it is useful to study the two-parameter model.

Setting 2 (Illustration with two-parameter model)**.** Let $\theta = (\theta_1, \theta_2)^\top$, and mutually independent observational and experimental estimates $\tilde{\theta}_j^{\text{obs}} - \theta_j \sim \mathcal{N}(b_j,\ \sigma_j^2)$, $\tilde{\theta}_j^{\text{exp}} - \theta_j \sim \mathcal{N}(0,\ v_j^2)$, $j = 1, 2$. For simplicity σ_j^2, v_j^2 are fixed. We may run one experiment due to budget constraints: $\mathcal{E} \in \{\{1\}, \{2\}\}$. If we pick j, we estimate $\hat{\theta}_j = \gamma_j\, \tilde{\theta}_j^{\text{exp}} + (1-\gamma_j)\, \tilde{\theta}_j^{\text{obs}}$, $\hat{\theta}_{-j} = \tilde{\theta}_{-j}^{\text{obs}}$, and, to first order, $\tau(\theta) - \tau(\hat{\theta}) = \omega_1(\theta_1 - \hat{\theta}_1) + \omega_2(\theta_2 - \hat{\theta}_2)$.

How do the variance and bias regret look like in this easier scenario? We can write

$$\alpha^\star = \min_{k\in\{1,2\}} \left\{\ \omega_{-k}^2\sigma_{-k}^2\ +\ \omega_k^2 \frac{\sigma_k^2\, v_k^2}{\sigma_k^2 + v_k^2}\ \right\}, \qquad \beta^\star = \min_{k\in\{1,2\}} \omega_k^2.$$

Here, $\alpha^\star$ follows from the variance-only optimal shrinkage $\gamma_k = \sigma_k^2/(\sigma_k^2 + v_k^2)$ (i.e., the oracle choice when $B = 0$); $\beta^\star$ equals the smallest between the two sensitivity parameters ω. Both depend on the choice of the estimator *and* on the class of experiments.

We cannot achieve both $\alpha^\star$ and $\beta^\star$, since these require different choices of γ. Instead,

$$\alpha(j,\gamma_j) \equiv \underbrace{\omega_{-j}^2\sigma_{-j}^2}_{\text{var obs estimate}} + \underbrace{\omega_j^2\big[(1-\gamma_j)^2\sigma_j^2 + \gamma_j^2 v_j^2\big]}_{\text{var exp and obs estimate}}, \quad \beta(j,\gamma_j) \equiv \Big(\ \underbrace{|\omega_{-j}|}_{\substack{\text{bias obs}\\ \text{estimate}/B}} + \underbrace{(1-\gamma_j)|\omega_j|}_{\substack{\text{bias obs minus}\\ \text{bias exp}/B}}\Big)^2.$$

The very first step is optimizing over the estimator (assuming here for simplicity γ can take any value). The following characterizes the regret optimal $\gamma^\star$ (see Appendix B.1).

$$\gamma_j^\star\ =\ \begin{cases} 1, & \text{if } \dfrac{\alpha(j,\gamma_j)}{\alpha^\star} < \dfrac{\beta(j,\gamma_j)}{\beta^\star} \text{ for all } \gamma_j \in \big(\tfrac{\sigma_j^2}{\sigma_j^2+v_j^2},\, 1\big), \\ \dfrac{\sigma_j^2}{\sigma_j^2+v_j^2}, & \text{if } \dfrac{\alpha(j,\gamma_j)}{\alpha^\star} > \dfrac{\beta(j,\gamma_j)}{\beta^\star} \text{ for all } \gamma_j \in \big(\tfrac{\sigma_j^2}{\sigma_j^2+v_j^2},\, 1\big), \\ \text{the unique } \gamma_j \in \big(\tfrac{\sigma_j^2}{\sigma_j^2+v_j^2},\, 1\big) \text{ s.t. } \dfrac{\alpha(j,\gamma_j)}{\alpha^\star} = \dfrac{\beta(j,\gamma_j)}{\beta^\star}, & \text{otherwise.} \end{cases} \tag{14}$$

At the boundaries, if the variance regret $\alpha/\alpha^\star$ is always smaller than the bias regret $\beta/\beta^\star$, the optimal weight is $\gamma_j^\star = 1$ (i.e., use only the experimental estimate). If instead the variance regret always dominates, the optimal choice is the variance-optimal $\gamma_j^\star = \sigma_j^2/(\sigma_j^2 + v_j^2)$.

When the solution is an interior solution for γ, the optimal design minimizes both the bias and variance regret (now equalized), so that

$$j^\star \in \arg\min_{j\in\{1,2\}} \underbrace{\big\{\omega_{-j}^2\sigma_{-j}^2 + \omega_j^2\big[(1-\gamma_j^\star)^2\sigma_j^2 + (\gamma_j^\star)^2 v_j^2\big]\big\}}_{\text{overall variance at } \gamma_j^\star} = \arg\min_{j\in\{1,2\}} \underbrace{\big\{|\omega_{-j}| + |1-\gamma_j^\star||\omega_j|\big\}}_{\text{worst–case bias}/|B| \text{ at } \gamma_j^\star}.$$

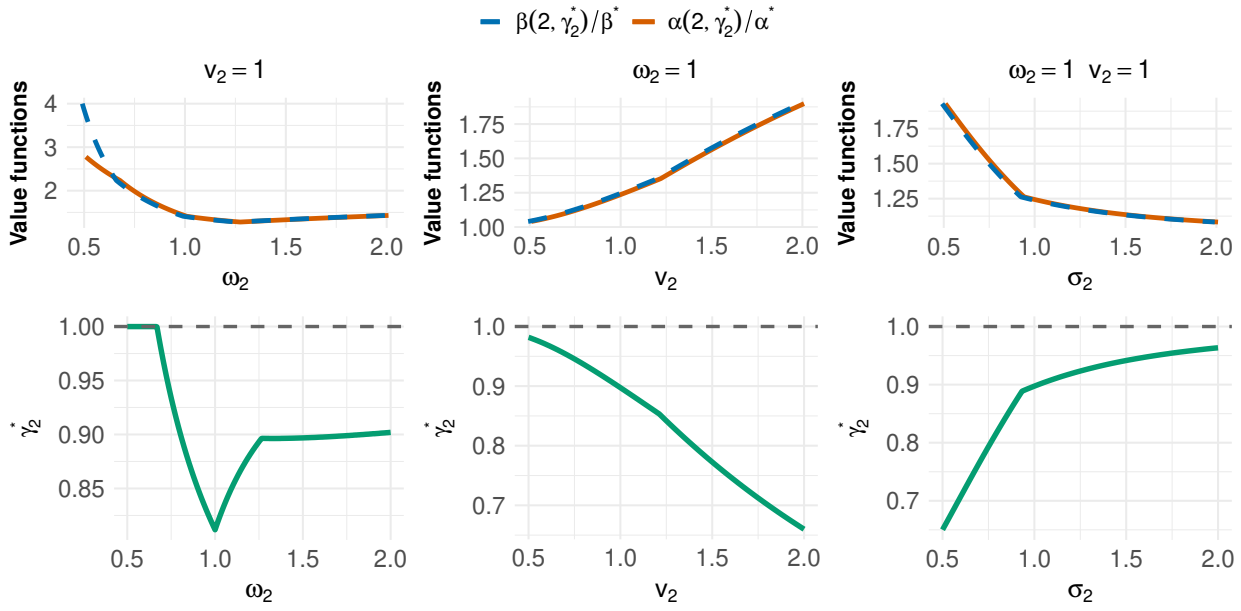

Figure 2: Example for $j = 2$. *Top row:* relative bias $\beta/\beta^\star$ (dashed) and relative variance $\alpha/\alpha^\star$ (solid), both evaluated at the optimal weight $\gamma_2^\star$. *Bottom row:* the optimal weight $\gamma_2^\star$. Columns vary, respectively, (a) ω_2 with $v_1 = 0.5$, $v_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $\omega_1 = 1$; (b) v_2 with $\omega = (0.9, 1)$, $v_1 = 1$, $\sigma = (1, 1)$; (c) σ_2 with $\omega = (0.9, 1)$, $v = (1, 1)$, $\sigma_1 = 1$.

However, cases at the extreme boundaries can also occur. For instance, if the variance regret $\alpha/\alpha^\star$ or the bias regret $\beta/\beta^\star$ uniformly dominates the other for one of the experiments, the optimal design minimizes the dominating term. These are desiderable properties as the optimal design always prioritizes the dominating term. Appendix Table 5 provides a complete overview of all possible scenarios. More complex scenarios where researchers may allocate sample sizes to each experiment are possible, see Section 5.

Example 4 (Numerical illustration)**.** For an illustrative example, consider choosing $\gamma_2^\star$ when experiment $j = 2$ is twice as precise as $j = 1$, with $\sigma_1^2 = \sigma_2^2 = 1$, $v_1 = 1$, $v_2 = 0.5$, and $\omega_1 = 1$. The optimal $\gamma_2^\star$ depends not only on the features of arm 2 but also on those of the alternative arm -2, $(v_{-j}^2, \sigma_{-j}^2, \omega_{-j})$, through the oracle benchmarks $(\alpha^\star, \beta^\star)$. Figure 2 illustrates this by plotting $\alpha(2, \gamma_2^\star)/\alpha^\star$ and $\beta(2, \gamma_2^\star)/\beta^\star$ at the optimal weight (top row) and the corresponding $\gamma_2^\star$ (bottom row), as we vary ω_2, v_2, and σ_2 column by column. Varying ω_2 yields two key effects: both very small and very large ω_2 push $\gamma_2^\star$ toward one, but for different reasons. When ω_2 is small, the oracle bias declines faster than the experimental variance (since $\beta^\star = \omega_2$), while when ω_2 is large, the oracle bias is $\beta^\star = \omega_1$ and ω_2 affects only the feasible experiment's bias. Consequently, both regret and the optimal estimator depend on the oracle's choice of *which* experiment it would run.

In the interior region for ω_2, the optimal weight first moves toward the variance-optimal

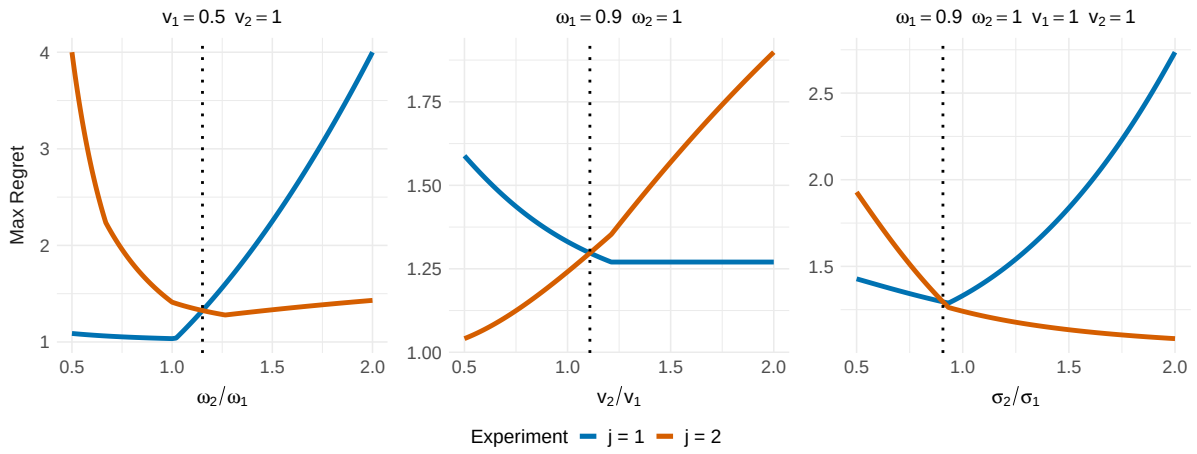

Figure 3: Example for $j = 2$ (regret). Different colors correspond to the regret for choosing either experiment. Figure reports $\max\{\alpha(j, \gamma_j^\star)/\alpha^\star, \beta(j, \gamma_j^\star)/\beta^\star\}$ for $j \in \{1, 2\}$ as the x–axis parameter varies (columns: ω_2, v_2, σ_2). The vertical dashed line marks indifference, where the two curves intersect; to its left/right, the optimal experiment is the one with lower max regret. Columns vary, respectively, (a) ω_2 with $v_1 = 0.5$, $v_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $\omega_1 = 1$; (b) v_2 with $\omega = (0.9, 1)$, $v_1 = 1$, $\sigma = (1, 1)$; (c) σ_2 with $\omega = (0.9, 1)$, $v = (1, 1)$, $\sigma_1 = 1$.

solution (as the oracle bias falls) and then rises again toward a bias-preferred solution. When we vary the experimental and observational variances, we obtain the expected comparative statics: $\gamma_2^\star$ decreases with the experimental variance and increases with the observational variance. These patterns translate into the regret of each arm (Figure 3). For arm $j = 2$, regret initially falls sharply as ω_2 rises and then declines more gradually once the bias component is already small. For arm $j = 1$, regret is at best slightly decreasing for small changes in ω_2 (when variance effects dominate) but eventually increases with ω_2, because the oracle increasingly favors arm 2 and the opportunity cost of selecting arm 1 rises. Table 4 summarizes these patterns; further discussion is provided in Appendix B.2. □

4 General framework

4.1 Moment selection

In many scenarios, we may expect that experiments only identify relevant moments in the data. We now extend the framework to generalized method of moments (GMM) estimators. In addition, we allow for a more general class of ambiguity sets and multivalued estimands.

Consider a vector of moment conditions $g(\theta) \in \mathbb{R}^p$ that stacks all moments (observational and experimental) potentially available for estimation. Let Λ denote the Jacobian of the moment function with respect to θ, evaluated at θ (i.e., $\Lambda := \partial g(\theta)/\partial \theta^\top$). For simplicity, we

will assume that Λ is known (i.e., moments are linear in θ). This simplifies exposition but it is not necessary in the local asymptotic framework we present in Section 4.3, where Λ can be replaced by the Jacobian of $g(\theta)$ evaluated at the observational estimates. For a given covariance matrix Σ of the sample moment vector, let $\bar{g}_\Sigma$ denote the sample analogue with

$$\mathbb{E}[\bar{g}_\Sigma] = b, \qquad \mathbb{V}(\bar{g}_\Sigma) = \Sigma,$$

where the vector b captures the biases in the moments. The researcher may decide to access only some of the entries of $\bar{g}$ by selecting a subset of experiments due to budget constraints.

General misspecification set. Let $\mathcal{I} \subseteq \{1, \ldots, p\}$ index all potential experimental moments researchers may select (unbiased by design) and $\mathcal{I}^c$ the observational ones (potentially misspecified). We posit bounded misspecification under arbitrary norms

$$\mathcal{B}_l(B) \;=\; \Big\{ b \in \mathbb{R}^p : \; b_j = 0 \text{ for } j \in \mathcal{I}, \;\; ||b_{j\in\mathcal{I}^c}||_l \leq B \text{ for } j \in \mathcal{I}^c \Big\}, \tag{15}$$

where the ambiguity set is defined relative to an arbitrary norm $||\cdot||_l$. Examples include the $\ell_1, \ell_2, \ell_\infty$ norms as special cases, as well as norms that may weight observational biases differently (see Example 6). The choice of the norm can be arbitrary assuming it is not data-dependent.

Estimator and design. Given the empirical moments $\bar{g}_\Sigma$, we consider (asymptotically) linear estimators of the form

$$\hat{\theta}(W, \Sigma) - \theta \;=\; \Gamma_\Lambda(W)\, \bar{g}_\Sigma, \quad \Gamma_\Lambda(W) \equiv -(\Lambda^\top W \Lambda)^{-1} \Lambda^\top W \tag{16}$$

where W and Σ are chosen by the researcher from a feasible set. The structure of the estimator follows a standard GMM asymptotic expansion (e.g. Andrews et al., 2020).[5]

Here, W may have columns of (effectively) zeros, which excludes the corresponding moments from the estimator. This occurs when researchers faces a trade-off in which experiment to run, leading to different moments $\bar{g}$ observed in the experiment.

The matrix Σ depends on the experimental design (e.g., sample–size allocations across experimental moments). The constraints on W and Σ effectively limits the class of experiments, sample size allocations and potentially estimators the researcher may consider.

[5]The weighting matrix W may be singular. The expansion $\hat{\theta} - \theta \doteq -(\Lambda^\top W \Lambda)^{-1} \Lambda^\top W\, \bar{g}_\Sigma$ holds whenever $\Lambda^\top W \Lambda$ is nonsingular, which can be forced through the set of constraints.

Assumption 3 (Constraint set)**.** Let $(W, \Sigma) \in \mathcal{D}'$ for a set $\mathcal{D}'$ so that $\Lambda^\top W \Lambda$ is invertible, $\Gamma_\Lambda(W)\Sigma\Gamma_\Lambda(W)^\top$ has uniformly bounded entries and is strictly positive definite.

Assumption 3 requires that the variance of the *selected* (reweighted) moments is uniformly bounded and strictly positive definite (even when W does not select some moments). This ensures that any subset of moments the researcher chooses to use is nondegenerate.

Example 5 (GMM moment selection)**.** Let $\theta \in \mathbb{R}^2$ and

$$g(\theta) = \left(g_1^{\text{obs}}(\theta),\, g_2^{\text{obs}}(\theta),\, g_1^{\text{exp}}(\theta),\, g_2^{\text{exp}}(\theta)\right)^\top \in \mathbb{R}^4,$$

where the first two components are observational and the last two are experimental. Suppose at most one experimental moment can be used. Selection is encoded by the GMM weight matrix $W \in \mathbb{R}^{4\times 4}$. A simple class of weighting matrices is

$$W^{(1)} = \begin{pmatrix} w_{\text{obs},1} & \rho & 0 & 0 \\ \rho & w_{\text{obs},2} & 0 & 0 \\ 0 & 0 & w_{\text{exp},1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad W^{(2)} = \begin{pmatrix} w_{\text{obs},1} & \rho & 0 & 0 \\ \rho & w_{\text{obs},2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{\text{exp},2} \end{pmatrix},$$

where $W^{(1)}$ selects the second experimental moment g_1^{exp} and sets the weight on g_2^{exp} to zero, while $W^{(2)}$ selects g_2^{exp} and excludes g_1^{exp}. In both cases, the observational block $(g_1^{\text{obs}}, g_2^{\text{obs}})$ is retained. These matrices are admissible provided that $\Lambda^\top W \Lambda$ is invertible. □

Target and MSE. Let the estimand be postentially multivalued $\tau(\theta) \in \mathbb{R}^q$ as $\tau(\theta) = \Omega\theta$ with $\Omega \in \mathbb{R}^{q\times d}$ with Ω different from the zero matrix. Our objective is the sum of MSE for each entry of τ, that is equal to

$$\text{MSE}_b(W, \Sigma) = ||\Omega\Gamma_\Lambda(W)\, b||_2^2 \;+\; \text{Trace}\left(\Omega\Gamma_\Lambda(W)\,\Sigma\,\Gamma_\Lambda(W)^\top\Omega^\top\right). \tag{17}$$

The key distinction now is that the variance and bias regret depend on Ω *and* Λ.

Definition 3 (Moment based variance and bias regret)**.** For a given matrix M, denote $||M||_{l,*} = \sup_{u:||u||_l\leq 1} ||Mu||_2$. In addition denote $[M]_{.,\mathcal{I}^c}$ the matrix A after removing the columns whose indexes are not in $\mathcal{I}^c$. With a slight abuse of notation, let

$$\alpha_\Omega(W, \Sigma) \;\equiv\; \text{Trace}\left(\Omega\Gamma_\Lambda(W)\,\Sigma\,\Gamma_\Lambda(W)^\top\Omega^\top\right), \qquad \beta_{l,\Omega}(W) \;\equiv\; \left\|[\Omega\Gamma_\Lambda(W)]_{.,\mathcal{I}^c}\right\|_{l,*}^2.$$

Let $\alpha_\Omega^\star, \beta_{l,\Omega}^\star$ their corresponding minimizers over $(W, \Sigma) \in \mathcal{D}'$. The moment-based variance and bias regret are defined as $\alpha_\Omega/\alpha_\Omega^\star, \beta_{l,\Omega}/\beta_{l,\Omega}^\star$ respectively.

The bias now depends on the interaction between the sensitivity Ω, the moment Jacobian Λ and the identity of which moments are observational and therefore potentially biased. For unidimensional $\tau(\theta)$ the bias corresponds to the squared dual norm of $\Omega\Gamma_\Lambda(W)$ with respect to $||\cdot||_l$. The weighting matrix W determines how these moments influence the downstream estimator $\hat{\theta}$.

As before, (and with a slight abuse of notation) we consider the adaptation regret of the following form:

$$\mathcal{R}_l(W,\Sigma) \equiv \sup_{B\geq 0} \frac{\sup_{b\in\mathcal{B}_l(B)} \mathrm{MSE}_b(W,\Sigma)}{\inf_{(W',\Sigma')\in\mathcal{D}'} \mathrm{MSE}_b(W',\Sigma')}. \tag{18}$$

The next theorem generalizes Theorem 1 to this more general framework.

Theorem 2 (Regret characterization)**.** *Let Assumptions 2, 3 hold, $\mathcal{R}$ be as defined in Equation* (18)*,with $\alpha,\beta,\alpha^\star,\beta^\star$ as in Definition 3. Consider an estimator as in Equation* (16)*. Then*

$$\mathcal{R}_l(W,\Sigma) = \max\left\{ \frac{\alpha_\Omega(W,\Sigma)}{\alpha^\star_\Omega}, \frac{\beta_{l,\Omega}(W)}{\beta^\star_{l,\Omega}} \right\}. \tag{19}$$

Proof. See Appendix A.1. □

Theorem 2 shows how our framework generalizes to (i) choosing the moments and the estimator (through W); and (ii) arbitrary norms characterizing the ambiguity set. The key insight is to show that the regret is quasi-convex under (approximate) linearity in the moments relative to the dual norm of the ambiguity set.

Example 6 (Weighted worst-case bias)**.** Let $\tau(\theta)\in\mathbb{R}$ and consider an ambiguity set of the form $\tilde{\mathcal{B}}(B) \equiv \left\{ b\in\mathbb{R}^p : |b_j|\leq k_j B \text{ for all } j=1,\dots,p \right\}$, for known rescaling factors k (e.g., the standard deviation for different outcomes to measure each θ_j). Then the corresponding bias regret is $\tilde{\beta}/\tilde{\beta}^\star$ with $\tilde{\beta}(\mathcal{E},\gamma) \equiv \left(\sum_{j\notin E} k_j|\omega_j| + \sum_{j\in E} k_j|\omega_j|\left|1-\gamma_j\right| \right)^2$, $\tilde{\beta}^* \equiv \min_{\mathcal{E}\in\mathcal{S}} \left(\sum_{j\notin\mathcal{E}} k_j|\omega_j| \right)^2$, and the overall regret equals $\max\left\{ \frac{\alpha(W,\Sigma)}{\alpha^\star}, \frac{\tilde{\beta}(W)}{\tilde{\beta}^\star} \right\}$. □

4.2 Confidence sets and partially identified τ

Next, we generalize our framework to minimizing confidence interval length, allowing for a potentially partially identified $\tau(\theta)$. For simplicity, we let $\tau(\theta)\in\mathbb{R}$ be a scalar. Specifically, researchers may not know $\tau(\cdot)$ exactly and only know functions $\bar{\tau},\underline{\tau}$ with

$$\bar{\tau}(\theta)=\bar{\omega}^\top\theta, \qquad \underline{\tau}(\theta)=\underline{\omega}^\top\theta, \quad \tau(\theta)\in[\underline{\tau}(\theta),\bar{\tau}(\theta)]$$

where, as for $\tau(\theta)$, linearity can be justified as a first-order linear approximation for $\bar{\tau}(\theta)$ and $\underline{\tau}(\theta)$ under the local asymptotics in Section 4.3 for smooth functions $\bar{\tau}, \underline{\tau}$.[6] We implicitly assume that either $\bar{\omega}$ or $\underline{\omega}$ (or both) differ from the zero vector to avoid trivial solutions.

For a generic weight vector ω let $\alpha_\omega(W, \Sigma), \beta_{l,\omega}(W)$ as in Definition 3.

Confidence intervals under bias. For a candidate design (W, Σ) and a bias vector $b \in \mathbb{R}^p$, define two-sided $(1-\eta)$ lower and upper bounds for $\tau(\theta)$ by

$$\begin{aligned}
\ell_b(W, \Sigma) &\equiv \underline{\tau}(\hat{\theta}(W)) \;-\; \underline{\omega}^\top \Gamma_\Lambda(W) b \;-\; z_{1-\eta/2} \sqrt{\alpha_{\underline{\omega}}(W, \Sigma)}, \\
u_b(W, \Sigma) &\equiv \bar{\tau}(\hat{\theta}(W)) \;-\; \bar{\omega}^\top \Gamma_\Lambda(W) b \;+\; z_{1-\eta/2} \sqrt{\alpha_{\bar{\omega}}(W, \Sigma)},
\end{aligned}$$

where $z_{1-\eta/2}$ is the standard normal $(1-\eta/2)$ quantile and $\hat{\theta}(W)$ is the GMM estimator in (16). These bounds adjust for the bias $\omega^\top \Gamma_\Lambda(W) b$ in $\omega^\top \hat{\theta}(W)$ for each envelope separately.

An audience, indexed by a worst–case bias radius $B \geq 0$, forms a worst–case confidence interval (i.e., a bias-aware confidence interval a la Armstrong and Kolesár (2018)) by choosing the most conservative endpoints over b in the ambiguity set $\mathcal{B}_l(B)$ in (15):

$$L_{l,B}(W, \Sigma) \;\equiv\; \left[\inf_{b \in \mathcal{B}_l(B)} \ell_b(W, \Sigma), \; \sup_{b \in \mathcal{B}_l(B)} u_b(W, \Sigma) \right],$$

and we denote by $|L_{l,B}(W, \Sigma)|$ its total length.

The researcher does not need to know the audience's specific choice of B and can simply report the point estimates and associated standard errors; the audience can then form their own confidence set from these statistics. However, the researcher may care about the expected length of the confidence interval formed by the audience.

An oracle that knows B minimizes the worst-case expected length of the audience's confidence length relative to the size of the sharp identified set (worst-case under the audience's prior $b \in \mathcal{B}_l(B)$). Its (ex-ante) expected loss is defined as

$$\mathcal{L}_{l,B}(W, \Sigma) \;\equiv\; \sup_{b_0 \in \mathcal{B}_l(B)} \Big\{ \mathbb{E}_{W,\Sigma,b_0} \big[|L_{l,B}(W, \Sigma)| \big] - \big(\bar{\tau}(\theta) - \underline{\tau}(\theta) \big) \Big\}. \tag{20}$$

Here $\mathcal{L}_{l,B}(W, \Sigma)$ measures how informative the expected confidence interval is, relative to the length of the identified set $\bar{\tau}(\theta) - \underline{\tau}(\theta)$, worst-case over b_0 in $\mathcal{B}_l(B)$. Here, b_0 determines the expectation of $\hat{\theta}(W)$, and ultimately the length of the bias-aware identified set.

[6] When the envelope for $\tau(\theta)$ is non-smooth, we recommend replacing it with smooth upper and lower envelopes. In this case, following Section 4.3, $\bar{\omega}$ and $\underline{\omega}$ can be interpreted as the gradients of these smooth bounds evaluated at the observational estimates.

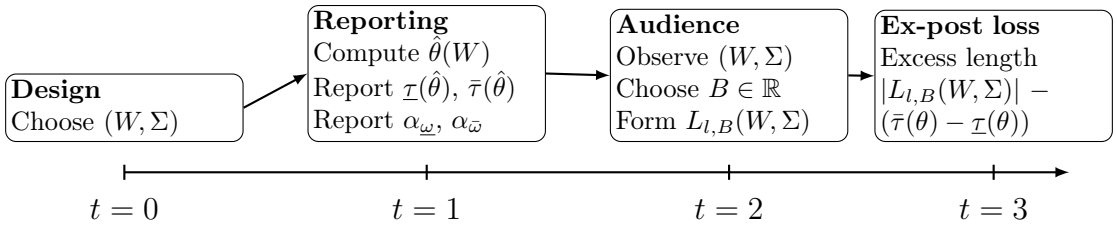

Figure 4: Timeline of design, reporting, audience confidence set, and loss.

As in previous sections, we define the corresponding regret as

$$\tilde{\mathcal{R}}_l(W, \Sigma) \equiv \sup_{B \geq 0} \frac{\mathcal{L}_{l,B}(W, \Sigma)}{\inf_{(W', \Sigma') \in \mathcal{D}'} \mathcal{L}_{l,B}(W', \Sigma')}.$$

Example 7 (Confidence set in a complete model)**.** Consider a complete model with $\bar{\tau}(\theta) = \underline{\tau}(\theta) = \tau(\theta)$ and $\bar{\omega} = \underline{\omega} = \omega$. Then we can show

$$\mathcal{L}_{l,B}(W, \Sigma) = 2\, z_{1-\eta/2} \sqrt{\alpha_\omega(W, \Sigma)} + 2\, B \sqrt{\beta_{l,\omega}(W)}.$$

In particular, when the parameter is point identified, the excess confidence interval length depends only on the variance index α_ω and the bias index $\beta_{l,\omega}$, corresponding to standard bias-aware confidence intervals (Armstrong and Kolesár, 2018).

Example 8 (Simple incomplete model)**.** Consider a randomized experiment with binary treatment $D \in \{0, 1\}$, outcome $Y \in [0, 1]$, and an indicator $R \in \{0, 1\}$ for whether Y is observed $(R = 1)$ or missing $(R = 0)$ from the survey. Let the treatment assignment probabilities be fixed by design, $\mathbb{P}(D = 1) = \pi_1$, and suppose $(Y(1), Y(0), R(1), R(0))$ are independent of D. The parameter of interest is the average treatment effect $\tau(\theta) \equiv \mathbb{E}\big[Y(1) - Y(0)\big]$, which is only partially identified without restrictions on the missingness process. Define the finite-dimensional parameter vector θ with $\theta_1 \equiv \mathbb{E}[YR1\{D = 1\}], \theta_2 \equiv \mathbb{P}(R = 0, D = 1), \theta_3 \equiv \mathbb{E}[YR1\{D = 0\}], \theta_4 \equiv \mathbb{P}(R = 0, D = 0)$. It follows that $\tau(\theta) \in \big[\underline{\tau}(\theta),\, \bar{\tau}(\theta)\big]$, where $\underline{\tau}(\theta) = \underline{\omega}^\top \theta, \bar{\tau}(\theta) = \bar{\omega}^\top \theta$, for $\underline{\omega} = \left(\frac{1}{\pi_1},\, 0,\, -\frac{1}{\pi_0},\, -\frac{1}{\pi_0}\right)^\top, \bar{\omega} = \left(\frac{1}{\pi_1},\, \frac{1}{\pi_1},\, -\frac{1}{\pi_0}\right)^\top$. Researchers choose π_1 in their experiment while using evidence from a country different from the one of the experimental population for observational estimates of θ. The loss captures loss in information of the worst-case estimated identified set relative to the sharp identified set. □

Regret optimal design (and estimator). Define

$$A(W, \Sigma) \equiv z_{1-\eta/2}\big(\sqrt{\alpha_{\bar{\omega}}(W, \Sigma)} + \sqrt{\alpha_{\underline{\omega}}(W, \Sigma)}\big), \quad C_l(W) \equiv \sqrt{\beta_{l,\bar{\omega}}(W)} + \sqrt{\beta_{l,\underline{\omega}}(W)} + \sqrt{\beta_{l,\bar{\omega}-\underline{\omega}}(W)},$$

and the corresponding oracle solutions

$$A^\star \equiv \inf_{(W,\Sigma)\in\mathcal{D}'} A(W,\Sigma), \qquad C_l^\star \equiv \inf_{(W,\Sigma)\in\mathcal{D}'} C_l(W).$$

The first component A depends on the variance of the estimated parameters (and envelopes) and the second component C_l depends on the bias. We now show how our results directly extend to this scenario.

Theorem 3 (Regret for confidence interval length)**.** *Let Assumptions 2 and 3 hold and consider an estimator as in* (16)*. Then, for any* $(W,\Sigma)\in\mathcal{D}'$*,*

$$\tilde{\mathcal{R}}_l(W,\Sigma) = \max\left\{\frac{A(W,\Sigma)}{A^\star},\ \frac{C_l(W)}{C_l^\star}\right\}.$$

Proof. See Appendix A.2. □

Theorem 3 shows that the confidence-interval-length regret can be decomposed into a variance component $A(W,\Sigma)$ and a bias component $C_l(W)$, each divided by its smallest feasible value. This result mimics Theorem 2 allowing for partially identified models. It provides a ready to use expression for experimental design (and moment selection).

The following corollary shows that whenever the model is point identified the solutions that minimize the regret relative to confidence interval length *coincides* with the MSE optimal solution we derived in Section 4.1.

Corollary 1 (Equivalence of MSE and confidence-interval length optimal solutions)**.** *Let Assumptions 2 and 3 hold. Let* $\alpha, \beta_l, \alpha^\star, \beta_l^\star$ *be as in Definition 3, and consider an estimator as in* (16)*. Suppose* $\bar{\omega} = \underline{\omega} = \omega$ *(complete model). Then, for any* (W,Σ)*,*

$$\arg\min_{(W,\Sigma)\in\mathcal{D}'} \mathcal{R}_l(W,\Sigma) \ = \ \arg\min_{(W,\Sigma)\in\mathcal{D}'} \tilde{\mathcal{R}}_l(W,\Sigma),$$

where $\mathcal{R}_l(W,\Sigma)$ *denotes the MSE-based regret in Equation* (18)*.*

Proof. See Appendix A.3. □

In summary, when $\tau(\theta)$ is point identified, minimizing the regret of the confidence interval length leads to the same optimal design as minimizing MSE-based regret. In partially identified models, Theorem 3 shows that the confidence interval length maintains the same bias–variance decomposition, up to appropriate redefinition of the indices to account for both endpoints of the identified set.

4.3 Local asymptotics with non linear target

Finally, we extend our framework for (i) nonlinear moment functions $g(\theta)$ used in GMM in Section 4.1 and (ii) a smooth, potentially nonlinear estimand $\tau(\theta)$. Although we focus on known mapping $\tau(\cdot)$, similar reasoning applies to Section 4.2.

Linearization of the moments Assume the sample moments obey root–n scaling, i.e., $\mathbb{V}(\bar{g}_\Sigma) = O(n^{-1})$. Let $b = b_n$ be a sequence where retain the same misspecification structure as in Section 4.1: the moment vector has mean $b = b_n$ with $(b_n)_j = 0$ for $j \in \mathcal{I}$ (experiment-based moments) and $(b_n)_{\mathcal{I}^c}$ potentially nonzero. Our key assumption is a local condition on the moments bias of the form

$$\|b_n\|^2\sqrt{n} \;\to\; 0. \tag{21}$$

This condition builds on the local asymptotic frameworks in Andrews et al. (2020) and Armstrong and Kolesár (2021). In their context, $||b_n||^2 = 1/n$ matching the rate of convergence of the variance. In our case, b_n can grow faster, slower or at the same rate of the standard error (i.e., $b_n \propto n^{-\alpha}, \alpha > 1/4$), encompassing $||b_n||^2 = 1/n$ as a special case.

Under standard smoothness conditions, a GMM estimator with weighting matrix W satisfies

$$\sqrt{n}\big(\hat{\theta}(W,\Sigma) - \theta\big) = -(\Lambda^\top W \Lambda)^{-1}\Lambda^\top W\,\sqrt{n}\big(\bar{g}_\Sigma - b_n\big) + o_p(1),$$

so on the original (unscaled) scale the mean shift is $-(\Lambda^\top W\Lambda)^{-1}\Lambda^\top W\, b_n$ and the remainder is $o_p(n^{-1/2})$. The condition $\|b_n\|^2\sqrt{n} \to 0$ ensures that higher order terms are negligible under sufficient smoothness conditions, justifying the use of the linear expansion (16) for our design analysis. Because $b_n \to 0$, the Jacobian Λ can be consistently estimated by evaluating the derivative of $g(\theta)$ at a preliminary observational estimator $\tilde{\theta}^{\text{obs}} \to_p \theta$, assuming such estimates are available from an observational study (Andrews et al., 2020).

Linearization of the estimand. Let $\tau : \mathbb{R}^p \to \mathbb{R}$ be twice continuously differentiable near θ, with gradient $\omega(\theta) := \partial\tau(\theta)/\partial\theta$ and bounded Hessian. A Taylor expansion around θ yields

$$\tau(\hat{\theta}) - \tau(\theta) = \omega(\theta)^\top(\hat{\theta} - \theta) \;+\; \tfrac{1}{2}\,(\hat{\theta} - \theta)^\top H_\tau(\bar{\theta})\,(\hat{\theta} - \theta),$$

for some $\bar{\theta}$ between $\hat{\theta}$ and θ, and H_τ is the Hessian. It follows that we can write

$$\sqrt{n}\big(\tau(\hat{\theta}) - \tau(\theta)\big) = \omega(\theta)^\top\,\Gamma_\Lambda(W)\,\sqrt{n}\big(\bar{g}_\Sigma - b_n\big) \;+\; o_p(1),$$

because $\|\hat{\theta} - \theta\| = O_p(n^{-1/2} + \|b_n\|)$ and the local condition $\sqrt{n}\,\|b_n\|^2 \to 0$ renders the quadratic remainder $o_p(1)$ after $\sqrt{n}$ scaling.

Consequently, the regret expressions derived for the linear case continue to apply up to $o_p(1)$, with ω interpreted as the gradient $\omega(\theta)$ (or any consistent plug-in $\omega(\tilde{\theta}^{\text{obs}})$ from a preliminary estimator $\tilde{\theta}^{\text{obs}} \to_p \theta$).

Remark 3 (Rescaling and change of variables linking Example 3)**.** An equivalent formulation uses $\sqrt{n}$–rescaled coordinates (as in Example 3). Define the rescaled moment vector $\tilde{g} := \sqrt{n}\,\bar{g}$ and the rescaled estimation error $\tilde{\theta} \equiv \sqrt{n}\,(\hat{\theta} - \theta)$. Then

$$\mathbb{V}(\tilde{g}) = \Sigma = O(1), \qquad \mathbb{E}[\tilde{g}] = b_n \;\text{ with }\; b_n \equiv \sqrt{n}\,b,$$

so b_n is the asymptotic bias and Σ the asymptotic variance. Local misspecification in (21) is equivalent to $\frac{\|b_n\|^2}{\sqrt{n}} \to 0$, which coincides with Footnote 4 and Example 3. □

5 Empirical applications

In this section we provide two empirical applications, one for measuring GE effect and the second one for externally valid site selection.

5.1 An implementation guide for measuring GE effects

Combining experimental or pre-program observational data with structural models is increasingly used for program evaluation (Allcott et al., 2025; Attanasio et al., 2012; Meghir et al., 2022; Todd and Wolpin, 2006).[7] Because large-scale experimentation are often infeasible due to cost constraints (Muralidharan and Niehaus, 2017) these models are meant to extrapolate effects from small-scale experiments to learn general-equilibrium (GE) effects. This extrapolation typically relies on other experiments or observational data.

We present a (step-by-step) illustration below when these models are calibrated using previous studies conducted in a different country.

Question Consider a researcher evaluating a cash transfers for sending children to school in rural Kenya where such a program is not in place. Due to budget constraints, the researcher can only randomize at a small scale (partial equilibrium), but ultimately wishes to predict GE effects. Researchers have access to preliminary estimates from the Mexican PROGRESA experiment (Todd and Wolpin, 2006), which we may be concerned, lack external validity.

[7] Across all AEA articles in 2015-2025, about 10% of the experimental papers combine experimental results with structural models (i.e., one-third of the papers that use observational in combination with experimental estimates).

5.1.1 Step 1: model description

The first step for a researcher is to describe the modeling choices. We present a stylized model of school choice from Bonhomme and Weidner (2022) and Todd and Wolpin (2006).

Individual choices Let $S \in \{0,1\}$ denote school attendance, C consumption, Y (pre-transfer) household income, W the child's potential wage, and t the stipend when enrolled. Abstracting from covariates (we will introduce covariates in the estimation), utility is

$$U(C,S,t,\varepsilon) = \xi_1\, C + \xi_2\, CS + (\xi_3 - \xi_1 - \xi_2)\, tS + \xi_4\, S + S\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1), \tag{22}$$

with budget $C = Y + W(1-S) + tS$. The parametrization $(\xi_3 - \xi_1 - \xi_2)$ is without loss and simplifies expressions below. Enrollment satisfies $S = 1\big\{U(Y+t,1,t,\varepsilon) > U(Y+W,0,0,0)\big\}$. Letting $Z(Y,W,t) \equiv \xi_1 W - \xi_2 Y - \xi_3 t - \xi_4$, we have $P(S=1 \mid Y,W,t) = \Phi\big(-Z(Y,W,t)\big)$, with $\Phi(\cdot)$ denoting the standard Gaussian CDF.

General equilibrium effects We are interested in the effect of a small stipend t to all eligible (poor) households in rural Kenya. GE feedback is allowed through income and wages: for functions $y(t), w(t)$ and mean-zero idiosyncratic income and wage shocks $\varepsilon_{Yi}, \varepsilon_{Wi}$, let

$$Y_i(t) = y(t) + \varepsilon_{Yi}, \qquad W_i(t) = w(t) + \varepsilon_{Wi}.$$

Let $y_0 \equiv \partial_t y(t)|_{t=0}$ and $w_0 \equiv \partial_t w(t)|_{t=0}$. Define $\phi_0 \equiv \mathbb{E}\big[\phi\big(-Z\big(Y(0),W(0),0\big)\big)\big]$. Our estimand of interest is the marginal effect of an increase in a small transfer t to all eligible households

$$\left.\frac{\partial\, \mathbb{E}[\Pr(S{=}1 \mid Y(t),W(t),t)]}{\partial t}\right|_{t=0} = \phi_0 \cdot \big(\,\xi_3 + \xi_2 y_0 - \xi_1 w_0\,\big), \tag{23}$$

which decomposes into the direct stipend effect $\phi_0\xi_3$ and the indirect GE effects via income and wages, $\phi_0\xi_2 y_0$ and $-\phi_0\xi_1 w_0$. Under simple market clearing for labor supply and demand, we can show[8]

$$w_0 = \frac{\phi_0\, \xi_3}{\phi_0\, \xi_1 - d}, \tag{24}$$

where d denotes the slope of the demand curve divided by total hours of work.

[8]This follows from the following: set $L_s(W,t) = L_d(W)$, the labor supply equals the labor demand, and differentiate at $t=0$: $\frac{\partial L_s}{\partial W} w_0 + \frac{\partial L_s}{\partial t} = \frac{\partial L_d}{\partial W} w_0$. Suppose we have a constant number of hours worked for working child, denoted as H, and any child not at school is working. From the probit index, $\partial_t \mathbb{E}[S \mid \cdot] = \phi(-Z)\,\xi_3$ and $\partial_W \mathbb{E}[S \mid \cdot] = -\phi(-Z)\,\xi_1$ at $t=0$. Hence $\partial L_s/\partial t = -\phi_0\xi_3 H$ and $\partial L_s/\partial W = \phi_0\xi_1 H$. Dividing the equation by H we obtain the desired result.

5.1.2 Step 2: parametrization and transparency on biases

Given the model, the following step for researchers is to define the parameters of interest. Here, we define the parameters of interest as

$$\theta_1 \equiv \phi_0 \xi_2, \qquad \theta_2 \equiv \phi_0 \xi_3, \qquad \theta_3 \equiv -\phi_0 \xi_1, \tag{25}$$

with corresponding estimand

$$\tau(\theta) = \underbrace{\theta_2}_{\text{Direct effect}} + \underbrace{y_0}_{\text{Income multiplier}} \underbrace{\theta_1}_{\text{Income effect}} + \underbrace{w_0(\theta)\,\theta_3}_{\text{Wage effect}}, \qquad w_0(\theta) = \frac{\theta_2}{-\theta_3 - d}, \tag{26}$$

where the explicit form of $w_0(\theta)$ follows from (24).

We treat d as known for simplicity, using meta-analyses on demand elasticities, and $y_0 = 1.5$ based on experimental estimates in Kenya from Egger et al. (2022).[9] Specifically, $d = 0.5\frac{1-S_0}{W_0}$, where S_0 and W_0 are baseline probability to go to school (assuming all children not in school go to work) and baseline wages calibrated to the pre-experimental data from PROGRESA and 0.5 obtained from meta-studies in Espey and Thilmany (2000).[10]

In this example, the primary concern is bias in the observational estimator $\tilde{\theta}^{\text{obs}}$ constructed as described below, abstracting from potential biases in y_0, d which, for simplicity, we assume to be second-order. When these biases are considered first-order, researchers should let y_0, d be additional parameters in θ. This difference illustrates a feature of our framework: the chosen parameterization ensures transparency about which bias concerns drive the design; these considerations can be collected in a pre-analysis plan.

5.1.3 Step 3: observational study estimates

Using experimental data from Mexico's PROGRESA program, we estimate $(\xi_1, \xi_2, \xi_3, \xi_4)$ via probit with standard controls (age, distance to school, eligibility, year, highest grade attained). Estimation is conducted separately by gender, and we will be focusing here on the effects on female students.

We pool treated and control observations to estimate the schooling effect ξ_4, controlling for the individual-level subsidy. In turn, we obtain observational estimates $\left(\tilde{\theta}_1^{\text{obs}}, \tilde{\theta}_2^{\text{obs}}, \tilde{\theta}_3^{\text{obs}}\right)$ that map to the model parametrization. Standard errors are constructed using the Delta method with clustering at the village level. Following Section 4.3, we then construct $\omega =$

[9] Egger et al. (2022) report a total income/consumption multiplier $M_{\text{tot}} = \partial_t \mathbb{E}\big[Y_{\text{pre}}(t) + t\big]|_{t=0} = 2.5$; therefore the target derivative of pre-transfer income is $y_0 = \partial_t \mathbb{E}\big[Y_{\text{pre}}(t)\big]|_{t=0} = M_{\text{tot}} - 1$.

[10] One could also calibrate S_0, W_0 to pre-experimental data in Kenya directly when available.

$\left.\frac{\partial\tau}{\partial\theta}\right|_{\theta=\tilde{\theta}^{\text{obs}}}$. Point estimates, estimated sensitivity ω, and standard errors are reported below.

Parameter	$\tilde{\theta}^{\text{obs}}$	ω	σ
θ_1 (Income)	$5.42e^{-5}$	1.50	$6.56e^{-5}$
θ_2 (Subsidy)	$1.93e^{-3}$	1.98	$9.86e^{-4}$
θ_3 (Wage)	-1.85e^{-3}	-2.03	$1.01e^{-3}$

Observational estimates for female students, corresponding sensitivity parameter ω, and standard errors.

$$\Sigma^{\text{obs}} = \begin{bmatrix} 4.31 & -11.31 & 5.57 \\ -11.31 & 973.11 & -126.16 \\ 5.57 & -126.16 & 1038.56 \end{bmatrix} \times e^{-9}$$

Variance–covariance matrix of observational estimates for female students.

We observe heterogeneity in both the sensitivity weights ω the precision of the observational estimates.

5.1.4 Step 4: describing the candidate experiments

The following step is to list the feasible experiments, subject to budget constraints.

In this example, we consider a researcher who can afford only small, partial-equilibrium experiments in Kenya, and examine three possible experiments for a given sample size n:

- $j = \{1\}$: *Unconditional transfer (income shock).* The researcher randomizes a small income shock to a small fraction of households (implying no general-equilibrium effects). Under a first-order (Taylor) approximation, the experiment identifies

$$\left.\frac{\partial\,\mathbb{E}[\Pr(S{=}1 \mid Y, W, 0)]}{\partial Y}\right|_{t=0} = \theta_1,$$

  with precision v_1^2/n. The researcher optimizes over γ_1, while θ_2 and θ_3 are calibrated to the PROGRESA study.

- $j = \{2\}$: *Conditional cash transfer (stipend).* The researcher randomizes a small stipend t, conditional on attending school, to a small fraction of households (with prices held fixed). Under a first-order approximation, this identifies

$$\left.\frac{\partial\,\mathbb{E}[\Pr(S{=}1 \mid Y, W, t)]}{\partial t}\right|_{t=0} = \theta_2,$$

  with precision v_2^2/n. The researcher optimizes over γ_2, while θ_1 and θ_3 are calibrated to the PROGRESA study.

- $j = \{1, 2\}$: *Two-arm design.* The researcher runs (j=1) and (j=2) on independent samples, identifying (θ_1, θ_2) with precisions $(v_1^2/n_1,\ v_2^2/n_2)$ where $n_1 + n_2 = n$. They optimize over both γ_1 and γ_2, while θ_3 is calibrated to the PROGRESA study.

We consider two scenarios. In the first, the researcher (and oracle) can choose one or both experiments. In the second, they can choose only one of the two (either $j = \{1\}$ or $j = \{2\}$). Regret is defined accordingly, conditional on the set of feasible designs.

For simplicity, we calibrate v_1^2 as $\sigma_1^2 \times n^{\text{obs}}$, where σ_1^2 is the variance of $\tilde{\theta}_1^{\text{obs}}$ and n^{obs} is the observational-sample size ($n^{\text{obs}} = 1089$ for the female sample using data from eligible students in Todd and Wolpin (2006)). We calibrate v_2^2 analogously as $\sigma_2^2 \times n^{\text{obs}}$. In practice, researchers may use pilot studies to calibrate the experimental variances when available.

Experiment type	Identified parameter	Choice Variables	Calibrated standard error	Description
$j = \{1\}$	θ_1	γ_1	$6.56e^{-5}/\sqrt{n}$	*Unconditional transfer (income shock, UCT).* Randomize an income shock to a small fraction of households.
$j = \{2\}$	θ_2	γ_2	$9.86e^{-4}/\sqrt{n}$	*Conditional cash transfer (education shock, CCT).* Randomize a conditional transfer to a small fraction of households.
$j = \{1, 2\}$	(θ_1, θ_2)	$\gamma_1,\ \gamma_2,\ n_1$	$6.56e^{-5}/\sqrt{n_1}$ (arm 1), $9.86e^{-4}/\sqrt{n - n_1}$ (arm 2)	*Two-arm design.* Run (j=1) and (j=2) on independent samples with sample sizes respectively $n_1, n - n_1$.

Table 1: Design options, identified parameters, choice variables, and per-unit variances.

5.1.5 Step 5: experimental design

The final step is to design the experiment and report it in a pre-analysis plan.

Choice of the experiment when only one treatment arm is available Figure 6 (right panel) displays the selected arm as a function of n (together with the corresponding $\gamma_j^\star$). We encode "not chosen" as $\gamma_j^\star = 0$. For small samples ($n \lesssim 500$), the optimal choice is UCT: the experimental CCT estimator would be too noisy, so the resulting variance under CCT dominates, while the bias advantage of CCT over UCT is limited because ω_1 and ω_2 are of similar magnitude. Once n crosses $n \approx 500$, CCT becomes dominant: by combining observational and experimental estimates for CCT, we can achieve comparable variance, and the worst-case bias is lower than under UCT.

Choice of the estimator In addition to sample allocation, our method optimizes the shrinkage weights of the parameters of interest. Figure 6 (left panel) shows that when both arms can be run, we place essentially full weight on the experimental estimator for UCT

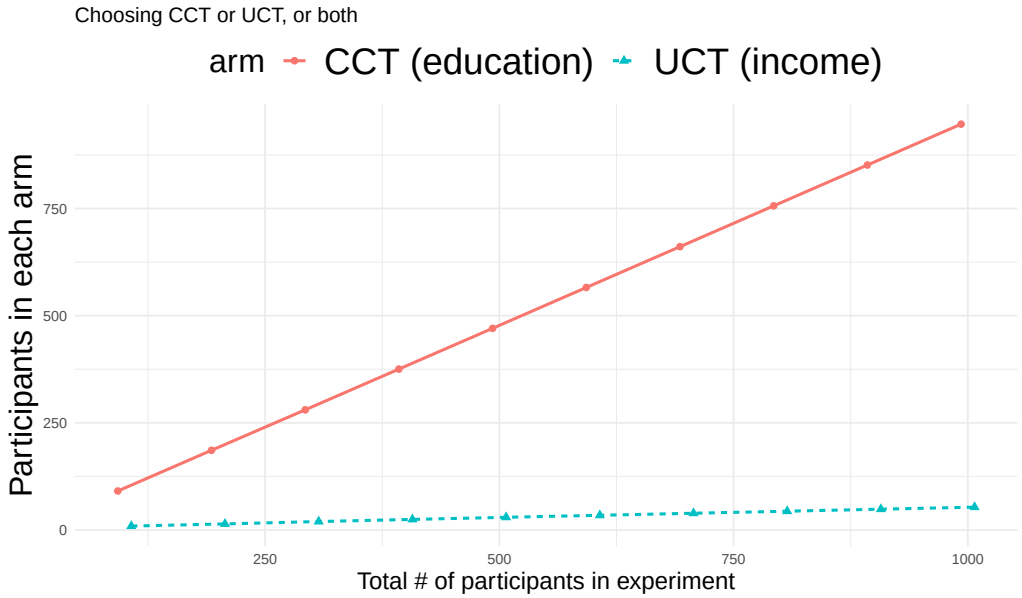

Figure 5: Regret–optimal sample allocation across treatment arms when both the unconditional cash transfer (UCT; income shock) and the conditional cash transfer (CCT; stipend) arms are feasible. The optimal solution is interior: the vast majority of participants (over 90%) are assigned to CCT, with a small but nonzero fraction assigned to UCT to hedge misspecification. This pattern reflects the higher payoff to learning about CCT in this application relative to UCT.

($\gamma^\star_{\text{UCT}} \approx 1$ across n), while for CCT we gradually shift toward the experimental estimator as precision improves: $\gamma^\star_{\text{CCT}}$ rises from about 0.5 at small n to approximately 0.9 by $n = 1000$. Intuitively, higher n lowers experimental variance, making it optimal to rely more heavily on experimental evidence for the CCT parameter.

Sample size allocation when both arms are feasible Figure 5 reports the sample allocation implied by the regret-optimal design when *both* treatment arms are feasible. The optimal allocation is interior: most participants (over roughly 90%) are assigned to the CCT arm (education), with a small but nonzero fraction assigned to the UCT arm (income shock). Two forces rationalize this pattern. First, the CCT parameter is more misspecification-sensitive ($\omega_2 \approx 1.98$ vs. $\omega_1 \approx 1.5$), so generating experimental evidence on CCT has a larger payoff in bias reduction. Second, the income effect θ_1's variance is about an order of magnitude smaller than the stipend effect θ_2's variance in our PROGRESA calibration, lowering the marginal returns to learn about UCT relative to CCT. Hence, while the designer keeps some allocation on UCT, most of the sample is placed on CCT; our framework makes this bias–variance trade-off explicit.

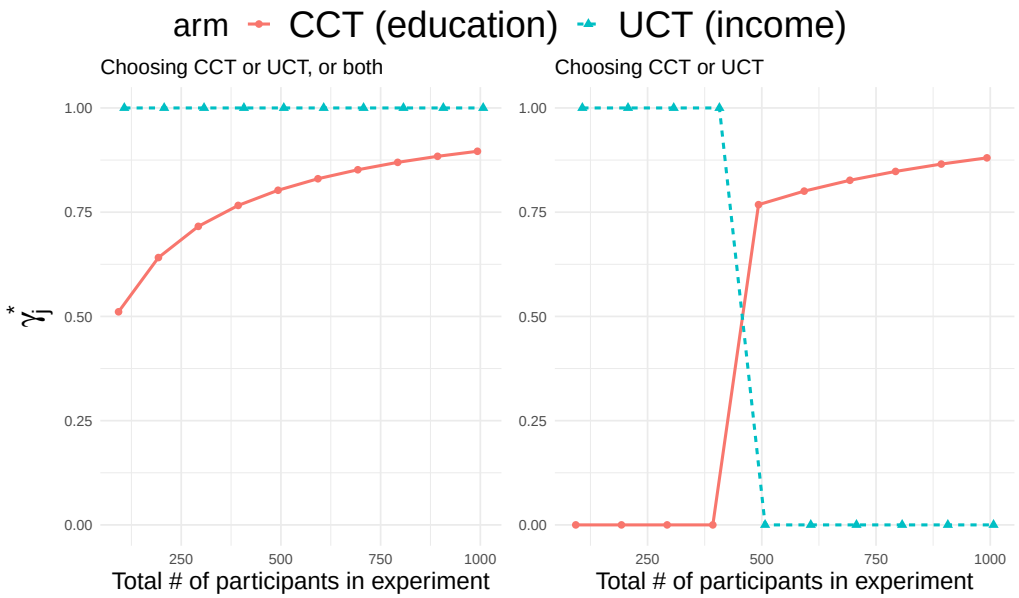

Figure 6: Optimal shrinkage weights and design choice as functions of the total sample size n. **Left panel:** When both arms are run, the weight on the UCT experimental estimator is essentially one across n, while the weight on the CCT experimental estimator, $\gamma^\star_{\text{CCT}}$, rises with precision (from about 0.5 at small n to roughly 0.9 by $n = 1000$), reflecting the increasing value of experimental evidence. **Right panel:** When restricted to a single arm, the optimal choice switches from UCT at small n (variance dominates under CCT) to CCT once n crosses a threshold (around $n \approx 500$), where bias considerations dominate and favor CCT. *Notes:* $\gamma_j^\star$ is the optimal weight on the experimental (vs. observational) estimator for parameter j; we set $\gamma_j^\star = 0$ to indicate that arm j is not chosen.

Regret comparisons Figure 7 reports worst-case regret for each design choice. The left panel shows that the regret of our method (green line, running both arms with an interior allocation) declines toward one as n increases, reflecting near-oracle performance under our normalization. In contrast, single-arm designs exhibit nonvanishing regret because bias remains first order even as variance shrinks. The right panel focuses on the single-arm problem: both the oracle and the researcher can only choose a single arm. In this case, UCT is preferable only for small n; beyond the same threshold (≈ 500), CCT yields uniformly lower regret as variance becomes second order relative to bias. The regret vanishes relative to the oracle that can only choose a single arm.

Implications When only one arm is feasible, CCT is preferable only once the total sample size is sufficiently large; otherwise UCT is preferred. This pattern reflects the interaction between bias and variance: for small samples, CCT's bias advantage is not large enough to compensate for its higher variance relative to UCT, whereas at larger samples the bias reduction becomes dominant. Existing designs based on variance considerations would fail

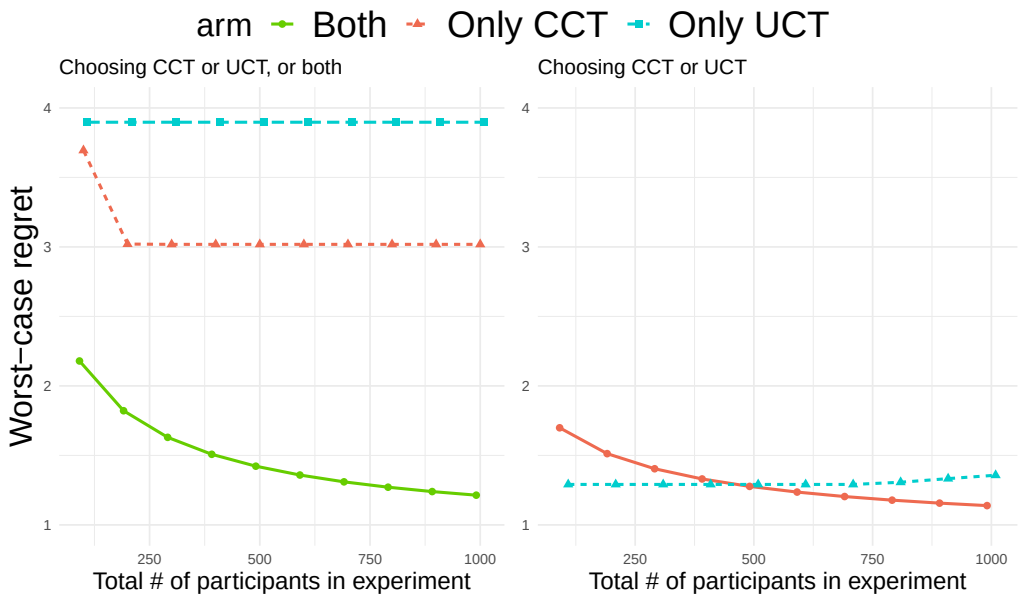

Figure 7: Worst–case regret by design. **Left panel:** Allowing both arms with an interior allocation yields regret that approaches one as n increases, indicating near–oracle performance under our normalization. In contrast, single–arm designs exhibit nonvanishing regret because bias remains first–order even as variance shrinks. **Right panel:** Focusing on the single–arm problem, UCT dominates only at small n; beyond the same threshold (≈ 500), CCT delivers uniformly lower regret as variance becomes second–order relative to bias. *Notes:* Regret is the worst–case objective normalized so that a value of one corresponds to the oracle benchmark; UCT = unconditional cash transfer, CCT = conditional cash transfer.

to provide solutions that balance bias and variance comparisons.

When feasible, it is instead optimal to run two arms with highly unbalanced sample, reflecting the relative contribution of the variance for each treatment arm.

Following these steps, researchers can similarly report their design in a pre-analysis plan.

5.2 Measuring the performance of our method for site selection

Next, we show how observational evidence can inform whom to recruit into an experiment. We consider the setting in Banerjee et al. (2024), who study how the expansion of microfinance affects village social networks in Karnataka, India. We focus on the first outcome reported by Banerjee et al. (2024) corresponding to the density of the network, i.e., the percentage of connections a random household in a village has relative to the village size.

Background In their observational analysis, the authors assemble panel data from 75 villages in Karnataka, 43 of which were exposed to microfinance. Because program rollout was not randomized across villages, they estimate effects using Difference-in-Differences (DiD).

While DiD is informative, the lack of randomized assignment may bias estimates in the presence of selection (e.g., Ghanem et al., 2022). Motivated by these limitations, the authors subsequently conducted an experimental evaluation in one metropolitan area, randomizing microfinance access across 104 urban neighborhoods. We only use observational variation for choosing the experimental design and estimator. We then use experimental variation generated by Banerjee et al. (2024) to validate our procedure.

Research question In practice, implementation costs often depend on how geographically dispersed the study sites are due to coordination and survey costs. Using preliminary observational estimates from Banerjee et al. (2024) in Karnataka, we ask: *Which area(s) in Karnataka should be prioritized for an experimental evaluation, and how many villages should be enrolled?*

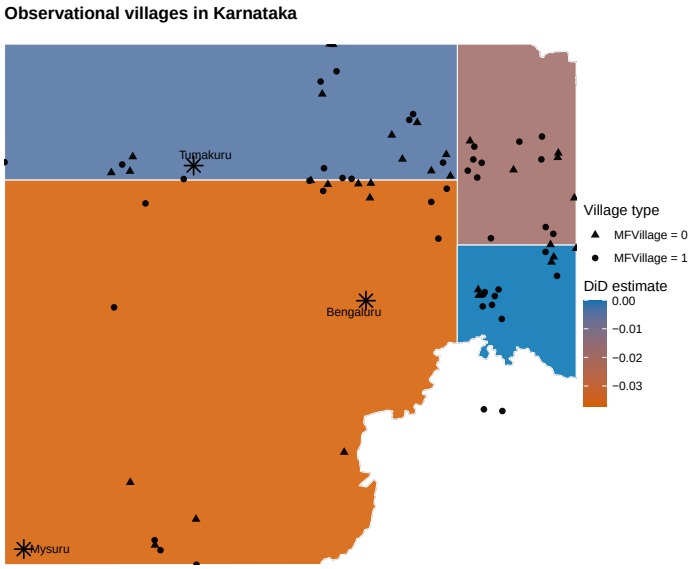

Figure 8: Observational villages and area-level DiD effects in Karnataka. The background heat map partitions the state into the four contiguous areas used to compute area-specific DiD estimates of microfinance's impact on network density; the color scale encodes the DiD value. Points mark village locations: circles denote villages exposed to microfinance and triangles denote unexposed villages in the observational sample (authors' survey). Major cities (Bengaluru, Mysuru, Tumakuru) are labeled for orientation.

Observational study We partition the observational sample into four geographically contiguous areas that group nearby villages; each area contains 11–12 villages that were exposed to microfinance during the study period. Areas are heterogeneous in terms of their overall population size. Figure 8 maps these four areas and reports the corresponding DiD point

estimates. Three of the four areas display negative estimated effects, with meaningful variation in magnitudes across areas. Table 2 summarizes, for each area, samples sizes variances, clustered at the village level, and population sizes (ω). These area-specific DiD estimates and their variances serve as our primary observational inputs for experimental design.[11] Using the 2011 population census, we compute ω as the population share of a given area.

Table 2: Area-level observational inputs for Karnataka villages. The outcome is network density (corresponding to the average share of connection of an individual to other individuals in a village). For each area, we report the numbers of treated (n_1) and untreated (n_0) villages, the pre-intervention variance of density pooled across arms (v^2_{pre}), the area-specific DiD estimate $\hat{\mu}$, and its sampling variance $\widehat{\sigma}^2$. Here ω denotes the relative population share obtained using the 2011 population census in India.

Area	n_1	n_0	v^2_{pre}	$\hat{\mu}$	$\widehat{\sigma}^2$	ω
1	11	6	0.000 457	−0.0187	0.000 045 3	0.252
2	12	9	0.001 75	−0.0377	0.000 140	0.246
3	11	12	0.001 38	−0.003 90	0.000 187	0.276
4	11	6	0.000 932	0.0148	0.000 130	0.227

Experimental design We consider a family of designs that (i) select $E \in \{1, \ldots, 4\}$ geographic areas in Karnataka from which to recruit experimental sites and (ii) assign n_1 villages to treatment and $n_0 = n_1$ to control (so the total sample is $n = 2n_1$). Here, $E = 1$ corresponds to recruiting from a single area, while $E = 4$ recruits from all four areas. We examine a grid of sample sizes that varies the number of participants from 10 up to 104 (52 treated units), with the latter corresponding to the size of the experiment in Banerjee et al. (2024). For variance calibration, we take $v^2_{\text{pre},a}, a \in \{1, \cdots, 4\}$ to be the pre-intervention, area-level variance of network density in Table 2. We assume that the variance of a single treated–control difference in the experiment is $2v^2_{\text{pre},a}$.

Experimental design We begin by studying which areas are selected and how sample is allocated when the total experimental sample is large ($n_1 = 52$) and the number of admissible areas may or may not be constrained. Figure 9 visualizes the resulting allocation across Karnataka's four observational areas for $E \in \{1, 2, 3, 4\}$. With $E = 1$, the algorithm concentrates recruitment in the Tumakuru area—the location with the highest sensitivity parameter (and second largest uncertainty in the observational estimates). As E relaxes to 2 and 3, the total sample splits across them in roughly (but not exactly) equal shares,

[11] We use DiD estimates to mimic the estimator used by the authors. An analogous analysis can be conducted after empirical-Bayes shrinkage of area-level estimates, omitted for brevity.

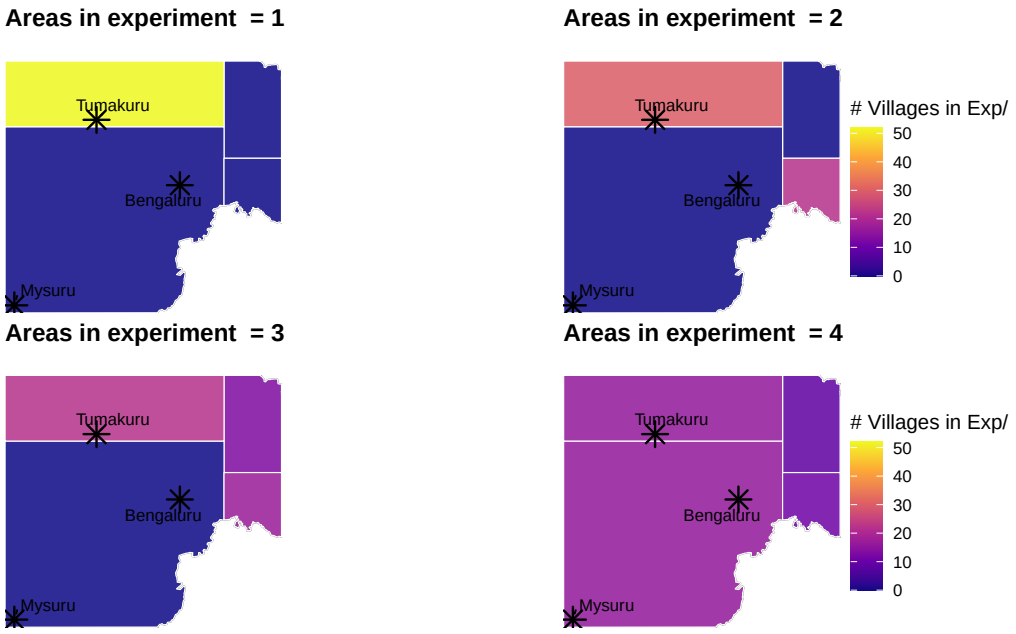

Figure 9: Optimal area selection and village allocation for a large experiment ($n_1 = 52$) under constraints on the number of areas E. Each panel corresponds to a value of E; shading indicates the total number of recruited villages in each area. With $E = 1$ the design concentrates in the area with the noisiest observational estimate; as E increases, allocation spreads across areas, reflecting a trade-off between observational uncertainty and experimental variance.

reflecting a trade-off between (i) exploiting heterogeneity in the observational variance and (ii) keeping experimental variance low. When $E = 4$, all areas are eligible and the allocation smooths further across space. In the rest of our discussion, we discuss properties when not all areas may be selected ($E < 4$).

Figure 14 tracks how area-level allocations evolve with the total number of villages. Appendix Figure 13 reports the optimal shrinkage $\gamma_j^\star$ by area as total villages vary. The solution is interior for all $E = 2$, and at the boundaries otherwise, with experimental estimates getting most of the weights: $\gamma_j^\star$ rises with sample size, approaching one from below as experimental noise diminishes.

MSE comparisons We compare three strategies: (i) our chosen design which chooses the area, sample size and $\gamma^\star$ as described in Section 3, (ii) a standard benchmark that selects areas uniformly at random, allocates villages evenly across the chosen areas, and sets $\gamma_j = 1$ for experimental estimates, and (iii) an oracle that knows the bias vector b and optimizes both design (areas and allocation) and γ.

Because b is unknown in practice, for this exercise, we calibrate it using the difference

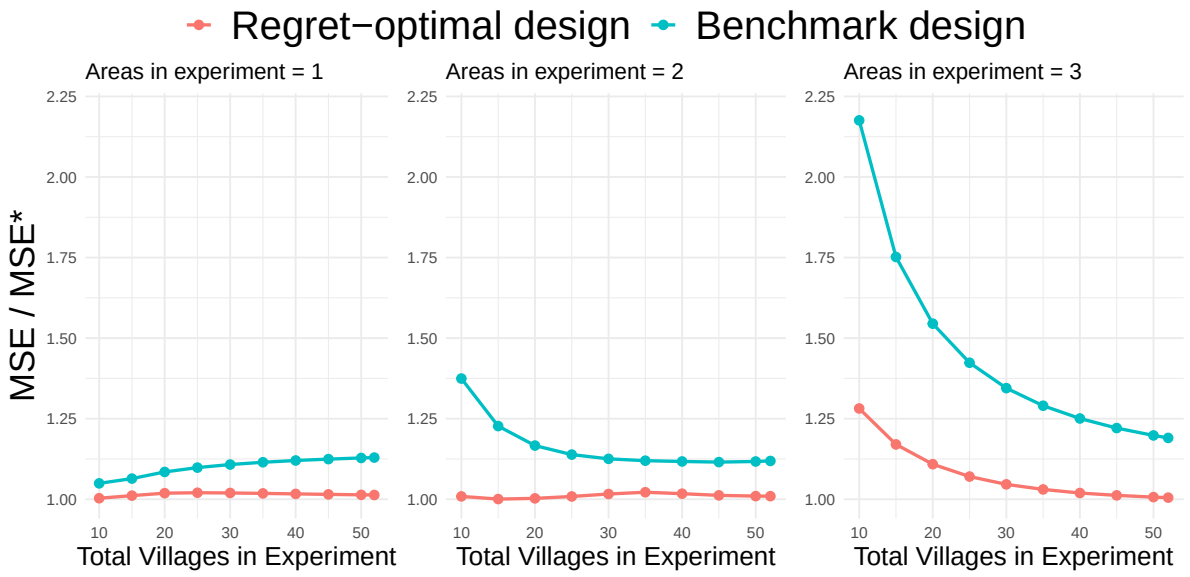

Figure 10: Relative MSE across designs under an hypothetical scenario where the bias equal the difference between the experimental and observational ATE estimate from Banerjee et al. (2024) as n_1 varies (from ten to fifty-two villages). The figure reports MSE/MSE* for (i) the proposed design (area targeting and optimal $\gamma^\star$), (ii) a random benchmark (uniform area selection and equal allocation), where MSE* denotes the MSE of the oracle that knows the bias, as n_1 varies and for each E. The proposed design closely tracks the oracle across n_1 and E, while the random benchmark remains substantially above, especially when E is small.

between the Hyderabad RCT ATE obtained from the follow-up experiment of Banerjee et al. (2024) and the Karnataka DiD average effect across sites, treating b as common across areas.

Figure 10 plots the ratio MSE/MSE*, where MSE* is the oracle's MSE. As the number of treated villages n_1 increases, all designs improve; because the denominator also falls with n_1, the ratio need not be monotone. Across $E \in \{1, 2, 3\}$, the random benchmark remains well above the oracle—by roughly 3–10 percentage points (pp) for $E = 1$, more than 10 pp for $E = 2$, and around 20 pp for $E = 3$—whereas our design tracks the oracle closely (within a few percentage points) even without knowledge of b. The gap is the largest when a larger number of areas E can be included, because our procedure can better emulate the oracle by selecting (and differentially weighting) favorable areas.

This result illustrates the merit of the method that improves MSE up to 20 percentage points compared to a standard benchmark and leads to a near-oracle solution.

6 Implications for practice

This paper studies experimental design in the presence of observational evidence. A key challenge is that the bias of the observational estimators is unknown in practice. We adopt a minimax proportional regret criterion that compares the mean-squared error (MSE) of a candidate design to that of an oracle that knows the worst-case bias. This reveals a fundamental trade-off between precision and robustness. The optimal design balances the design's variance normalized by the smallest achievable variance (variance gap) and its worst-case bias normalized by the smallest attainable bias (bias gap). We propose a procedure that jointly determines: (i) how to combine observational and experimental evidence; (ii) how to allocate precision across experiments given budget constraints; and (iii) which treatment arm and/or sub-population to include in the experiment with fixed experimental costs.

In practice, the workflow is:

- **Define the estimand(s) of interest.** Specify $\tau(\theta)$ for a known mapping τ and unknown parameters $\theta \in \mathbb{R}^p$. For example, $\tau(\theta)$ may represent a counterfactual, a general-equilibrium effect, or an average impact across locations. Adopt a parametrization in which some (but not necessarily all) components of θ can be learned experimentally; this clarifies what the experiment can identify. When such parametrization is not readily available, Section 4.1 shows how our procedure applies to τ identified through arbitrary moment restrictions and Section 4.2 when it is partially identified.

- **Assemble informative observational evidence.** Collect observational estimates $\tilde{\theta}^{\text{obs}}$ and their covariance $\tilde{\Sigma}^{\text{obs}}$. These serve as informative (but potentially biased) baselines. Such evidence may come from observational designs, structural estimates with pre-experimental data, or prior experiments conducted in different contexts.

- **Compute the sensitivity parameters.** Using the observational baseline, compute the sensitivity weights $\omega = \frac{\partial \tau(\theta)}{\partial \theta}\big|_{\tilde{\theta}^{\text{obs}}}$, which quantify how bias in each coordinate of θ propagates to $\tau(\theta)$. Large $|\omega_j|$ indicates greater payoff to learning the jth component.

- **Specify feasibility constraints and calibrate experimental variance.** Enumerate the admissible design set, and the budget on sample size allocation. Calibrate per-unit experimental variances using pilot studies or historical data to form the set of feasible designs $\mathcal{D}$.

- **Run the method and decide.** Optimize over the design and corresponding estimators. The output is a pre-analysis plan with the selected arm(s), sample sizes, and

pre-specified combination rule (γ), and optimization is conducted via a mixed-integer quadratic program in Algorithm 1.

The applicability of our framework spans a wide range of settings in economics and beyond. Examples include estimating general equilibrium effects or structural models (Attanasio et al., 2012; de Albuquerque et al., 2025; Kreindler et al., 2023; Meghir et al., 2022; Todd and Wolpin, 2006); choosing among alternative treatment arms in factorial designs (Bandiera et al., 2025; Muralidharan et al., 2020); and deciding where to run the next experiment for external validity (Gechter et al., 2024; Olea et al., 2024). In industrial organization, applications include choices between demand-side and supply-side interventions (Bergquist and Dinerstein, 2020), the effects of information acquisition in markets (Allende et al., 2019; Larroucau et al., 2024), and decisions about which additional data source to acquire to improve statistical analysis (Allcott et al., 2025). Beyond economics, medical applications include allocating sample size across subgroups and allocating doses across treatment arms (e.g., Manski, 2025; Morita et al., 2017; Porter et al., 2024).

Several open questions remain for future research. These include settings where researchers have well-specified priors about bias such as from Bayesian models (Gechter et al., 2024), or other modeling choices (Olea et al., 2024). This may be potentially incorporated through additional moment conditions as in Section 4.1. Additional future extensions include sequential or adaptive experimental choices (e.g. Cesa-Bianchi et al., 2025).

References

Abadie, A. and J. Zhao (2021). Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196*.

Allcott, H., J. C. Castillo, M. Gentzkow, L. Musolff, and T. Salz (2025). Sources of market power in web search: Evidence from a field experiment. Technical report, National Bureau of Economic Research.

Allende, C., F. Gallego, and C. Neilson (2019). Approximating the equilibrium effects of informed school choice. Technical report.

Andrews, I., N. Barahona, M. Gentzkow, A. Rambachan, and J. M. Shapiro (2025). Structural estimation under misspecification: Theory and implications for practice. *The Quarterly Journal of Economics*, qjaf018.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2020). Transparency in structural research. *Journal of Business & Economic Statistics 38*(4), 711–722.

Andrews, I. and J. M. Shapiro (2021). A model of scientific communication. *Econometrica 89*(5), 2117–2142.

Armstrong, T. B., P. Kline, and L. Sun (2024). Adapting to misspecification.

Armstrong, T. B. and M. Kolesár (2018). Optimal inference in a class of regression models. *Econometrica 86*(2), 655–683.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

Athey, S., R. Chetty, and G. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S., R. Chetty, and G. Imbens (2025). The experimental selection correction estimator: Using experiments to remove biases in observational estimates. Technical report, National Bureau of Economic Research.

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

Atkinson, A. C. and V. Fedorov (1975). The design of experiments for discriminating between two rival models. *Biometrika 62*(1), 57–70.

Attanasio, O. P., C. Meghir, and A. Santiago (2012). Education choices in mexico: using a structural model and a randomized experiment to evaluate progresa. *The Review of Economic Studies 79*(1), 37–66.

Bai, Y. (2019). Optimality of matched-pair designs in randomized controlled trials. *Available at SSRN 3483834*.

Bandiera, O., A. Jalal, and N. Roussille (2025). The illusion of time: Gender gaps in job search and employment. Technical report, National Bureau of Economic Research.

Banerjee, A., E. Breza, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, and C. Kinnan (2024). Changes in social network structure in response to exposure to formal credit markets. *Review of Economic Studies 91*(3), 1331–1372.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–1230.

Bassi, V., M. E. Kahn, N. L. Gracia, T. Porzio, and J. Sorin (2022). Jobs in the smog: Firm location and workers' exposure to pollution in african cities. Technical report, National Bureau of Economic Research.

Bergquist, L. F. and M. Dinerstein (2020). Competition and entry in agricultural markets: Experimental evidence from kenya. *American Economic Review 110*(12), 3705–3747.

Bertsimas, D., M. Johnson, and N. Kallus (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research 63*(4), 868–876.

Bhattacharya, D. (2013). Evaluating treatment protocols using data combination. *Journal of Econometrics 173*(2), 160–174.

Bonhomme, S. and M. Weidner (2022). Minimizing sensitivity to model misspecification. *Quantitative Economics 13*(3), 907–954.

Box, G. E. and N. R. Draper (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association 54*(287), 622–654.

Breza, E., A. G. Chandrasekhar, and D. Viviano (2025). Generalizability with ignorance in mind: learning what we do (not) know for archetypes discovery. *arXiv preprint arXiv:2501.13355*.

Cesa-Bianchi, N., R. Colomboni, and M. Kasy (2025). Adaptive maximization of social welfare. *Econometrica 93*(3), 1073–1104.

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical science*, 273–304.

Chaudhuri, P. and P. A. Mykland (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association 88*(422), 538–546.

Chetty, R., N. Hendren, and L. F. Katz (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review 106*(4), 855–902.

Christensen, T. and B. Connault (2023). Counterfactual sensitivity and robustness. *Econometrica 91*(1), 263–298.

Cytrynbaum, M. (2021). Optimal stratification of survey experiments. *arXiv preprint arXiv:2111.08157*.

de Albuquerque, A., F. Finan, A. Jha, L. Karpuska, and F. Trebbi (2025). Decoupling taste-based versus statistical discrimination in elections. Technical report, National Bureau of Economic Research.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Empirical mse minimization to estimate a scalar parameter. *arXiv preprint arXiv:2006.14667*.

Dominitz, J. and C. F. Manski (2017). More data or better data? a statistical decision problem. *The Review of Economic Studies 84*(4), 1583–1605.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics 22*(1), 238–270.

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

Dutz, D., I. Huitfeldt, S. Lacouture, M. Mogstad, A. Torgovitsky, and W. Van Dijk (2021). Selection in surveys: Using randomized incentives to detect and account for nonresponse bias. Technical report, National Bureau of Economic Research.

Egger, D., J. Haushofer, E. Miguel, P. Niehaus, and M. Walker (2022). General equilibrium effects of cash transfers: Experimental evidence from Kenya. *Econometrica 90*(6), 2603–2643.

Espey, M. and D. D. Thilmany (2000). Farm labor demand: A meta-regression analysis of wage elasticities. *Journal of Agricultural and Resource Economics*, 252–266.

Gechter, M. (2022). Combining experimental and observational studies in meta-analysis: A debiasing approach. Working paper, Pennylvania State University and London School of Economics.

Gechter, M., K. Hirano, J. Lee, M. Mahmud, O. Mondal, J. Morduch, S. Ravindran, and A. S. Shonchoy (2024). Selecting experimental sites for external validity. *arXiv preprint arXiv:2405.13241*.

Gerber, A. S. and D. P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Ghanem, D., P. H. Sant'Anna, and K. Wüthrich (2022). Selection and parallel trends. *arXiv preprint arXiv:2203.09001*.

Higbee, S. D. (2024). Experimental design for policy choice.

Hu, Y., H. Zhu, E. Brunskil, and S. Wager (2024). Minimax-regret sample selection in randomized experiments. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 1209–1235.

James, W., C. Stein, et al. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379. University of California Press.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(1), 85–112.

Kallus, N., A. M. Puli, and U. Shalit (2018). Removing hidden confounding by experimental grounding. *Advances in neural information processing systems 31*.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*(3), 324–338.

Kasy, M. and A. Sautmann (2019). Adaptive treatment assignment in experiments for policy choice.

Katz, J. and H. Allcott (2025). Digital media mergers: Theory and application to facebook-instagram. Technical report, Working paper.

Kiefer, J. and J. Wolfowitz (1959). Optimum designs in regression problems. *The annals of mathematical statistics 30*(2), 271–294.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kreindler, G., A. Gaduh, T. Graff, R. Hanna, and B. A. Olken (2023). Optimal public transportation networks: Evidence from the world's largest bus rapid transit system in jakarta. Technical report, National Bureau of Economic Research.

Larroucau, T., I. Rios, A. Fabre, and C. Neilson (2024). College application mistakes and the design of information policies at scale. *Unpublished paper, Arizona State University, Tempe*.

List, J. A., S. Sadoff, and M. Wagner (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics 14*(4), 439–457.

López-Fidalgo, J., C. Tommasi, and P. C. Trandafir (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 69*(2), 231–242.

Manski, C. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Manski, C. F. (2025). Using limited trial evidence to credibly choose treatment dosage when efficacy and adverse effects weakly increase with dose. *Epidemiology 36*(1), 60–65.

Manski, C. F. and A. Tetenov (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference 137*(6), 1998–2010.

Manski, C. F. and A. Tetenov (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences 113*(38), 10518–10523.

Meghir, C., A. M. Mobarak, C. Mommaerts, and M. Morten (2022). Migration and informal insurance: Evidence from a randomized controlled trial and a structural model. *The Review of Economic Studies 89*(1), 452–480.

Montiel Olea, J., C. Qiu, and J. Stoye (2023). Decision theory for treatment choice with partial identification. *Preprint*.

Morita, S., P. F. Thall, and K. Takeda (2017). A simulation study of methods for selecting subgroup-specific doses in phase 1 trials. *Pharmaceutical statistics 16*(2), 143–156.

Muralidharan, K. and P. Niehaus (2017). Experimentation at scale. *Journal of Economic Perspectives 31*(4), 103–24.

Muralidharan, K., M. Romero, and K. Wüthrich (2020). Factorial designs, model selection, and (incorrect) inference in randomized experiments. NBER Working Paper.

Olea, J. L. M., B. Prallon, C. Qiu, J. Stoye, and Y. Sun (2024). Externally valid selection of experimental sites via the k-median problem. *arXiv preprint arXiv:2408.09187*.

Porter, S., T. A. Murray, and A. Eaton (2024). Phase i/ii design for selecting subgroup-specific optimal biological doses for prespecified subgroups. *Statistics in Medicine 43*(28), 5401–5411.

Rambachan, A., R. Singh, and D. Viviano (2024). Program evaluation with remotely sensed outcomes. *arXiv preprint arXiv:2411.10959*.

Reeves, S. W., S. Lubold, A. G. Chandrasekhar, and T. H. McCormick (2024). Model-based inference and experimental design for interference using partial network data. *arXiv preprint arXiv:2406.11940*.

Rosenman, E. T. and A. B. Owen (2021). Designing experiments informed by observational studies. *Journal of Causal Inference 9*(1), 147–171.

Rosenman, E. T., A. B. Owen, M. Baiocchi, and H. R. Banack (2022). Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine 41*(1), 65–86.

Russo, D. J., B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning 11*(1), 1–96.

Sacks, J. and D. Ylvisaker (1984). Some model robust designs in regression. *The Annals of Statistics*, 1324–1348.

Silvey, S. (2013). *Optimal design: an introduction to the theory for parameter estimation*, Volume 1. Springer Science & Business Media.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*.

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review 96*(5), 1384–1417.

Tsirpitzi, R. E., F. Miller, and C.-F. Burman (2023). Robust optimal designs using a model misspecification term. *Metrika 86*(7), 781–804.

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the sobolev classes. *The Annals of Statistics 26*(6), 2420–2469.

Viviano, D. (2020). Experimental design under network interference. *arXiv preprint arXiv:2003.08421*.

Wiens, D. P. (1998). Minimax robust designs and weights for approximately specified regression models with heteroscedastic errors. *Journal of the American Statistical Association 93*(444), 1440–1450.

A Proofs of main results

A.1 Proof of Theorem 1, Theorem 2

Theorem 1 is a special case of Theorem 2 proved below. We will prove Theorem 2 taking the uncertainty set over arbitrary balls $\mathcal{B}_l(B) = \{b : ||b||_l \leq B\}$ and denote $||\cdot||_{l,*}$ as in the main text.

Step 1 (worst-case MSE). By definition of dual norm,

$$\sup_{b\in\mathcal{B}_l(B)} ||\Omega\Gamma_\Lambda(W)\, b||_2^2 = \Big(B \,\big\|\big[\Omega\Gamma_\Lambda(W)\big]_{\cdot,\mathcal{I}^c}\big\|_{l,*}\Big)^2 = B^2\,\beta_l(W).$$

Hence $\sup_{b\in\mathcal{B}_l(B)} \mathrm{MSE}_b(W,\Sigma) = \alpha(W,\Sigma) + B^2\beta_l(W)$.

Step 2 (oracle envelope and basic properties). Let $\delta(t) := \inf_{W,\Sigma}\{\alpha(W,\Sigma) + t\beta_l(W)\}$. Recall that $\beta_l(W)$ and $\alpha(W,\Sigma)$ are bounded from above and $\alpha(W,\Sigma) > 0$ by Assumption 3 (using the fact that Ω contains some non zero entry). Being the pointwise infimum of affine functions of t, δ is concave, nondecreasing, and finite for $t \geq 0$. Moreover,

$$\delta(0) = \alpha^\star, \qquad \lim_{t\to\infty} \frac{\delta(t)}{t} = \beta_l^\star.$$

Step 3 (quasi-convexity and boundary maximization). Fix (α, β_l). Consider

$$\phi(t) := \frac{\alpha + t\beta_l}{\delta(t)}, \qquad t \geq 0.$$

Because the numerator is affine and the denominator is positive and concave in t, ϕ is quasi-convex on $[0,\infty)$. Hence, on any compact interval $[0,T]$, $\max_{t\in[0,T]}\phi(t)$ is attained at the boundary $\{0,T\}$. Letting $T\to\infty$ and using the limits in Step 2,

$$\sup_{t\geq 0}\phi(t) = \max\left\{\frac{\alpha}{\delta(0)},\ \lim_{t\to\infty}\frac{\alpha+t\beta_l}{\delta(t)}\right\} = \max\left\{\frac{\alpha}{\alpha^\star},\ \frac{\beta_l}{\beta_l^\star}\right\},$$

with the stated conventions $0/0 = 1$.

A.2 Proof of Theorem 3

Step 1 (closed form of $\mathcal{L}_{l,B}(W,\Sigma)$). Fix (W,Σ) and $B \geq 0$. By definition,

$$\ell_b(W,\Sigma) = \underline{\tau}(\hat{\theta}(W)) - \underline{\omega}^\top\Gamma_\Lambda(W)b - z_{1-\eta/2}\sqrt{\alpha_{\underline{\omega}}(W,\Sigma)},$$

so

$$\inf_{\|b\|_l \le B} \ell_b(W,\Sigma) = \underline{\tau}(\hat{\theta}(W)) - z_{1-\eta/2}\sqrt{\alpha_{\underline{\omega}}(W,\Sigma)} - \sup_{\|b\|_l \le B} \underline{\omega}^\top \Gamma_\Lambda(W) b.$$

Since $b_j = 0$ for $j \in \mathcal{I}$ and $\|b_{\mathcal{I}^c}\|_l \le B$, the dual norm gives

$$\sup_{\|b\|_l \le B} \underline{\omega}^\top \Gamma_\Lambda(W) b = B\sqrt{\beta_{l,\underline{\omega}}(W)}.$$

Hence

$$\inf_{\|b\|_l \le B} \ell_b(W,\Sigma) = \underline{\tau}(\hat{\theta}(W)) - z_{1-\eta/2}\sqrt{\alpha_{\underline{\omega}}(W,\Sigma)} - B\sqrt{\beta_{l,\underline{\omega}}(W)}.$$

Similarly, for the upper bound,

$$\sup_{\|b\|_l \le B} u_b(W,\Sigma) = \bar{\tau}(\hat{\theta}(W)) + z_{1-\eta/2}\sqrt{\alpha_{\bar{\omega}}(W,\Sigma)} + B\sqrt{\beta_{l,\bar{\omega}}(W)}.$$

Therefore the worst-case interval has length

$$|L_{l,B}(W,\Sigma)| = \big(\bar{\tau}(\hat{\theta}(W)) - \underline{\tau}(\hat{\theta}(W))\big) + z_{1-\eta/2}\big(\sqrt{\alpha_{\bar{\omega}}(W,\Sigma)} + \sqrt{\alpha_{\underline{\omega}}(W,\Sigma)}\big) + B\big(\sqrt{\beta_{l,\bar{\omega}}(W)} + \sqrt{\beta_{l,\underline{\omega}}(W)}\big).$$

Let b_0 denote the true bias, with $\|b_0\|_l \le B$. Under (16),

$$\bar{\tau}(\hat{\theta}(W)) - \underline{\tau}(\hat{\theta}(W)) = (\bar{\omega} - \underline{\omega})^\top \theta + (\bar{\omega} - \underline{\omega})^\top \Gamma_\Lambda(W) \bar{g}_\Sigma,$$

so

$$\mathbb{E}_{W,\Sigma,b_0}\big[\bar{\tau}(\hat{\theta}(W)) - \underline{\tau}(\hat{\theta}(W))\big] = (\bar{\omega} - \underline{\omega})^\top \theta + (\bar{\omega} - \underline{\omega})^\top \Gamma_\Lambda(W) b_0.$$

Subtracting the identified set length $\bar{\tau}(\theta) - \underline{\tau}(\theta) = (\bar{\omega} - \underline{\omega})^\top \theta$, we obtain

$$\mathbb{E}_{W,\Sigma,b_0}\big[|L_{l,B}(W,\Sigma)|\big] - \big(\bar{\tau}(\theta) - \underline{\tau}(\theta)\big) =$$
$$(\bar{\omega} - \underline{\omega})^\top \Gamma_\Lambda(W) b_0 + z_{1-\eta/2}\big(\sqrt{\alpha_{\bar{\omega}}(W,\Sigma)} + \sqrt{\alpha_{\underline{\omega}}(W,\Sigma)}\big) + B\big(\sqrt{\beta_{l,\bar{\omega}}(W)} + \sqrt{\beta_{l,\underline{\omega}}(W)}\big).$$

Taking the worst case over $\|b_0\|_l \le B$ implies

$$\mathcal{L}_{l,B}(W,\Sigma) = z_{1-\eta/2}\big(\sqrt{\alpha_{\bar{\omega}}(W,\Sigma)} + \sqrt{\alpha_{\underline{\omega}}(W,\Sigma)}\big) + B\big(\sqrt{\beta_{l,\bar{\omega}}(W)} + \sqrt{\beta_{l,\underline{\omega}}(W)} + \sqrt{\beta_{l,\bar{\omega}-\underline{\omega}}(W)}\big),$$

which yields the linear form $\mathcal{L}_{l,B}(W,\Sigma) = A(W,\Sigma) + B\, C_l(W)$.

Step 2 (oracle envelope and basic properties). Define

$$\delta(t) \;\equiv\; \inf_{(W,\Sigma)\in\mathcal{D}'} \{A(W,\Sigma) + t\, C_l(W)\}, \qquad t \ge 0.$$

By Assumption 3 (because either or both $\bar{\omega}, \underline{\omega}$ are not zero vectors), $A(W,\Sigma)$ and $C_l(W)$

are bounded from above and $A(W,\Sigma) > 0$ for all $(W,\Sigma) \in \mathcal{D}'$, so $\delta(t)$ is finite and strictly positive for all $t \geq 0$. Being the pointwise infimum of affine functions of t, δ is concave and nondecreasing on $[0,\infty)$. Moreover,

$$\delta(0) = A^\star, \qquad \lim_{t\to\infty} \frac{\delta(t)}{t} = C_l^\star.$$

Step 3 (quasi-convexity and boundary maximization). Fix (W,Σ) and set $A \equiv A(W,\Sigma)$ and $C \equiv C_l(W)$. For $t \geq 0$ define

$$\phi(t) \;\equiv\; \frac{A + t\,C}{\delta(t)}.$$

The numerator is affine in t and the denominator is positive and concave, so ϕ is quasi-convex on $[0,\infty)$. Hence, on any compact interval $[0,T]$, $\max_{t\in[0,T]} \phi(t)$ is attained at the boundary $\{0,T\}$. Letting $T \to \infty$ and using the limits in Step 2,

$$\sup_{t\geq 0} \phi(t) = \max\left\{ \frac{A}{\delta(0)},\ \lim_{t\to\infty} \frac{A + t\,C}{\delta(t)} \right\} = \max\left\{ \frac{A(W,\Sigma)}{A^\star},\ \frac{C_l(W)}{C_l^\star} \right\}.$$

By the definition of $\tilde{\mathcal{R}}_l(W,\Sigma)$ and the identity $\mathcal{L}_{l,B}(W,\Sigma) = A(W,\Sigma) + BC_l(W)$, this yields

$$\tilde{\mathcal{R}}_l(W,\Sigma) = \sup_{B\geq 0} \frac{\mathcal{L}_{l,B}(W,\Sigma)}{\inf_{(W',\Sigma')\in\mathcal{D}'} \mathcal{L}_{l,B}(W',\Sigma')} = \max\left\{ \frac{A(W,\Sigma)}{A^\star},\ \frac{C_l(W)}{C_l^\star} \right\},$$

which proves the theorem. □

A.3 Proof of Corollary 1

Whenever $\bar{\omega} = \underline{\omega} = \omega$, we can write $A(W,\Sigma) = \alpha_\omega(W,\Sigma)$ and $C_l(W) = \beta_{l,\omega}(W)$. From Theorem 3

$$\tilde{\mathcal{R}}_l(W,\Sigma) \;=\; \max\left\{ \frac{\alpha_\omega(W,\Sigma)}{\alpha^\star},\ \ \frac{\beta_{l,\omega}(W,\Sigma)}{\beta_l^\star} \right\}^{1/2}.$$

Because $\tilde{\mathcal{R}}_l(W,\Sigma) = \mathcal{R}_l(W,\Sigma)^2$ the set of minimizer remains unchanged, completing the proof.

B Examples

B.1 Two parameters examples in Equation (14)

By Theorem 1, for fixed $(\mathcal{E}, \Sigma)$ the adaptation regret as a function of the shrinkage vector equals

$$\mathcal{R}(\gamma) \;=\; \max\Big\{\alpha(\mathcal{E}, \Sigma, \gamma)/\alpha^\star,\; \beta(\mathcal{E}, \gamma)/\beta^\star\Big\},$$

with $\alpha^\star, \beta^\star$ constants.

Step 1. With independence and two parameters,

$$\alpha(\mathcal{E}, \Sigma, \gamma) = C + \omega_j^2\Big(\gamma_j^2 v_j^2 + (1-\gamma_j)^2\sigma_j^2\Big),$$

where C does not depend on γ_j. Hence $\alpha(j, \gamma_j) \equiv \alpha(\mathcal{E}, \Sigma, \gamma)$ is a strictly convex quadratic in γ_j with unique minimizer at

$$\gamma_j^{\text{var}} = \frac{\sigma_j^2}{\sigma_j^2 + v_j^2}.$$

Moreover, $\alpha(j, \gamma_j)$ is strictly increasing on $\big(\gamma_j^{\text{var}}, 1\big)$ and strictly decreasing on $\big[0, \gamma_j^{\text{var}}\big)$.

Write the worst–case bias component holding the other coordinate fixed as

$$\beta(j, \gamma_j) \;=\; \Big(A_j + |\omega_j| \cdot |1-\gamma_j|\Big)^2,$$

where $A_j \geq 0$ collects all terms not involving γ_j (including the contribution from the other coordinate). On $[0, 1]$, $|1-\gamma_j| = 1-\gamma_j$, so $\beta(j, \gamma_j)$ is strictly decreasing in γ_j and convex with minimum at $\gamma_j = 1$. Therefore, any minimizer satisfies

$$\gamma_j^\star \in \big[\gamma_j^{\text{var}},\, 1\big].$$

Step 2. On $\big(\gamma_j^{\text{var}}, 1\big)$ the map $\gamma_j \mapsto \alpha(j, \gamma_j)/\alpha^\star$ is strictly increasing, while $\gamma_j \mapsto \beta(j, \gamma_j)/\beta^\star$ is strictly decreasing. Hence the function

$$f(\gamma_j) \equiv \max\Big\{\alpha(j, \gamma_j)/\alpha^\star,\; \beta(j, \gamma_j)/\beta^\star\Big\}, \qquad \gamma_j \in \big[\gamma_j^{\text{var}}, 1\big],$$

is minimized either (i) at a boundary point, or (ii) at the unique interior point where the two arguments are equal (by strict monotonicity, at most one intersection exists). Therefore,

- If $\alpha(j, \gamma_j)/\alpha^\star < \beta(j, \gamma_j)/\beta^\star$ for all $\gamma_j \in \big(\gamma_j^{\text{var}}, 1\big)$, then $f(\gamma_j) = \beta(j, \gamma_j)/\beta^\star$ on that interval. Since this term is strictly decreasing, the minimizer is the right boundary $\gamma_j^\star = 1$.

- If $\alpha(j,\gamma_j)/\alpha^\star > \beta(j,\gamma_j)/\beta^\star$ for all $\gamma_j \in \big(\gamma_j^{\text{var}}, 1\big)$, then $f(\gamma_j) = \alpha(j,\gamma_j)/\alpha^\star$ on that interval. Since this term is strictly increasing there, the minimizer is the left boundary $\gamma_j^\star = \gamma_j^{\text{var}} = \sigma_j^2/(\sigma_j^2 + v_j^2)$.
- Otherwise, by the intermediate value theorem and strict monotonicity of the two curves, there exists a unique $\gamma_j \in \big(\gamma_j^{\text{var}}, 1\big)$ such that $\frac{\alpha(j,\gamma_j)}{\alpha^\star} = \frac{\beta(j,\gamma_j)}{\beta^\star}$.

B.2 Numerical example

Here, we present a numerical example in Setting 2.

Choice of $\gamma^\star$ Figure 2 (top row) reports $\alpha(2,\gamma_2^\star)/\alpha^\star$ and $\beta(2,\gamma_2^\star)/\beta^\star$ evaluated at the optimal weight while varying, column by column, ω_2, v_2, and σ_2. In the bottom row, we plot the corresponding $\gamma_2^\star$.

Observation 1: small ω_2 pushes $\gamma_2^\star$ toward one. The first (left) column sets $v_1 = v_2/2$ (the first experiment is more precise) and $\omega_1 = 1$ (with $\sigma_1 = \sigma_2 = 1$). For small ω_2 (low sensitivity of the second coordinate to bias), the optimal choice is $\gamma_2^\star = 1$. The reason is twofold. First, the bias ratio $\beta(2,\gamma)/\beta^\star = \big(|\omega_1| + |1-\gamma||\omega_2|\big)^2/\beta^\star$ is normalized by $\beta^\star = \min\{|\omega_1|, |\omega_2|\}^2 = \omega_2^2$ when $\omega_2 < \omega_1$; thus even a small observational component $|1-\gamma| > 0$ makes the ratio increase substantially. Second, the variance ratio $\alpha(2,\gamma)/\alpha^\star$ is multiplied by ω_2^2 in its $j=2$ contribution, so its dependence on γ becomes negligible as $\omega_2 \downarrow 0$. Together these forces make the bias ratio dominant and push $\gamma_2^\star$ to the boundary at 1. The implication is that for parameters with low ω_2, we typically shrink $\hat{\theta}_2$ more toward the experimental estimate $\tilde{\theta}_2^{\text{exp}}$.

Observation 2: when ω_2 is of the same order as ω_1, an interior solution emerges. As ω_2 increases toward ω_1 (roughly $0.6\,\omega_1$ and above), the normalization ceases to penalize $j=2$ as harshly, and an interior solution appears: $\gamma_2^\star$ moves down from 1 toward the variance–only weight, trading off bias and variance. The first kink in the bottom-left panel reflects this regime change, where $\gamma_2^\star$ is selected to equalize bias and variance.

Observation 3: for $\omega_2 > 1$, $\gamma_2^\star$ increases again. At $\omega_2 = \omega_1$ we observe a second kink, since the worst–case bias normalizer switches to $\beta^\star = \omega_1^2$. This raises the bias cost of putting weight on the observational estimate; as ω_2 increases above 1, $\gamma_2^\star$ rises again. This pattern is driven by the oracle benchmark, which optimizes both the estimator and the design.

Observation 4: for $\omega_2 \gtrsim 1.25$, $\gamma_2^\star$ increases slowly and stays interior. Around $\omega_2 \approx 1.25$ we observe a third kink: the variance normalizer $\alpha^\star$ switches branches (from favoring $k=1$ to favoring $k=2$ in the variance–only comparison). The optimal $\gamma_2^\star$ remains interior but

grows more slowly, because increases in $\gamma_2^\star$ now have a larger relative impact on $\alpha/\alpha^\star$.[12]

Observation 5: $\gamma_2^\star$ increases with experimental precision and decreases with observational precision The second and third columns vary v_2 and σ_2, respectively. In both, we fix $v_1 = 1$ and take $\omega_1 = 0.9\,\omega_2$ (so here $\omega_1 = 0.9$, $\omega_2 = 1$), with $\sigma_1 = \sigma_2 = 1$ unless varied. Across $v_2, \sigma_2 \in [0.5, 2]$, the solution is interior. As $v_2 \downarrow 0.5$ or $\sigma_2 \uparrow 2$, the optimal weight $\gamma_2^\star$ increases toward 1, placing nearly all weight on the experimental estimate. Conversely, as v_2 becomes large or σ_2 becomes small, $\gamma_2^\star$ approaches its lower bound, the variance–minimizing weight $\sigma_2^2/(\sigma_2^2 + v_2^2)$. This is a typical (and desired) behavior of shrinkage estimators.

Table 3: Example: Regimes for $\gamma_2^\star$ as a function of ω_2 for example in Appendix Figure 2 with $\sigma_1^2 = \sigma_2^2 = 1, v_1 = 1, v_2 = 0.5, \omega_1 = 1$.

Regime	$\beta^\star$	$\alpha^\star$	**Solution regime ($\gamma_2^\star$)**	**Trend of $\gamma_2^\star$**
$\boldsymbol{\omega_2 \ll \omega_1}$	$\lvert\omega_2\rvert^2$	$\omega_2^2\sigma_2^2 + \omega_1^2\frac{\sigma_1^2 v_1^2}{\sigma_1^2 + v_1^2}$	*Bias–dominant*	*Constant* $\gamma_2^\star = 1$
$\boldsymbol{\omega_2 < \omega_1}$	$\lvert\omega_2\rvert^2$	$\omega_2^2\sigma_2^2 + \omega_1^2\frac{\sigma_1^2 v_1^2}{\sigma_1^2 + v_1^2}$	*Interior*	*Decrease* in ω_2
$\boldsymbol{\omega_2 > \omega_1}$	$\lvert\omega_1\rvert^2$	$\omega_2^2\sigma_2^2 + \omega_1^2\frac{\sigma_1^2 v_1^2}{\sigma_1^2 + v_1^2}$	*Interior*	*Increase* in ω_2
$\boldsymbol{\omega_2 \gg \omega_1}$	$\lvert\omega_1\rvert^2$	$\omega_1^2\sigma_1^2 + \omega_2^2\frac{\sigma_2^2 v_2^2}{\sigma_2^2 + v_2^2}$	*Interior*	*Slow increase* in ω_2

Choice of the design Figure 11 compares the maximum regret of choosing experiment $j = 1$ versus $j = 2$ in three scenarios. The vertical dashed line marks the value where the two curves intersect and the designer is indifferent between $j = 1$ and $j = 2$. In each column, we keep the same parameterizations as in Figure 2.

Observation 1: The regret for $j = 2$ decreases rapidly in ω_2 and then decreases more slowly. In the first column, we vary ω_2 with the first experiment $j = 1$ having a smaller experimental variance, $v_1 = v_2/2$. The panel shows that a small ω_2 makes the regret of choosing experiment 2 larger than that of choosing experiment 1. This is because a small ω_2 corresponds to a small bias from not choosing $j = 2$. As ω_2 increases, the max–regret curve for $j = 2$ declines until it reaches $\omega_2 \approx 1.25$; after this point, the regret curve for $j = 2$ rises slowly, since a further increase in ω_2 is associated with an increase in estimator variance.

Observation 2: The regret for $j = 1$ decreases slowly and then increases rapidly in ω_2. The regret curve for choosing $j = 1$ as we vary ω_2 (fixing $\omega_1 = 1$) first decreases slowly

[12] A further kink could occur if the solution transitioned into a pure variance–dominant regime (where the boundary weight is chosen), which does not arise for the range of ω_2 shown here.

and then increases. The reason is that for $\omega_2 < \omega_1$, $\gamma_1^\star$ is an interior solution and the oracle squared bias is $\beta^\star = \omega_2^2$. Therefore, an increase in ω_2^2 raises the oracle's bias. When $\omega_2 > \omega_1$, $\gamma_1^\star$ becomes a boundary solution and $\beta^\star = \omega_1^2$. In this case, a larger ω_2 increases the variance of the estimator while the oracle bias $\beta^\star$ remains constant in ω_2. As a result, it becomes more attractive for the analyst to run the experiment with $j = 2$.

Observation 3: The regret for $j = 2$ is monotonically increasing in v_2^2, and the opposite holds for $j = 1$. The second plot shows that the regret from choosing $j = 2$ increases monotonically with the experimental variance v_2^2, while the regret from choosing $j = 1$ decreases correspondingly. The kinks in the curves are driven by regime shifts in the oracle solutions $(\beta^\star, \alpha^\star)$.

Observation 4: The regret for $j = 2$ is monotonically decreasing in σ_2^2. The third plot shows that the regret for $j = 2$ decreases monotonically with the observational variance σ_2^2. This is expected, since a larger σ_2^2 makes choosing $j = 2$ more appealing relative to relying on observation.

Observation 5: The regret for $j = 1$ first decreases slowly and then increases rapidly in σ_2^2. As we vary σ_2^2, the regret for $j = 1$ is initially (very) slowly decreasing. The reason is that a larger σ_2^2 increases the oracle variance $\alpha^\star$ faster than the variance of the estimator for $j = 1$ (with $\gamma_1^\star \approx 1$). However, this behavior does not affect the optimal solution: for smaller values of σ_2^2, choosing $j = 1$ remains preferable to choosing $j = 2$. When σ_2^2 is larger than a tipping point, however, the regret of choosing $j = 1$ increases rapidly with σ_2^2, making the first experiment no longer preferable to the second. This aligns with the intuition that, as observational variance grows, we should favor conducting the experiment for $j = 2$.

In sum, these patterns show how our framework disentangles the competing forces across signal strength and observational/experimental noise, guiding the analyst to transparent design choices even in complex regimes. Table 4 summarizes the discussion.

Table 4: Example: Qualitative behavior of maximum regret as ω_2, v_2, or σ_2 increase (rows) in Figure 11. Entries summarize the direction and relative speed of change in the maximum regret for choosing experiment $j = 1$ or $j = 2$.

	Small regime $(\omega_2 < \omega_1)$		**Large regime** $(\omega_2 > \omega_1)$	
Trend	*regret* $j = 1$	*regret* $j = 2$	*regret* $j = 1$	*regret* $j = 2$
$\uparrow$ ω_2	slow decrease	fast decrease	fast increase	slow increase
$\uparrow$ v_2	decrease	increase	decrease	increase
$\uparrow$ σ_2	slow decrease	decrease	fast increase	decrease

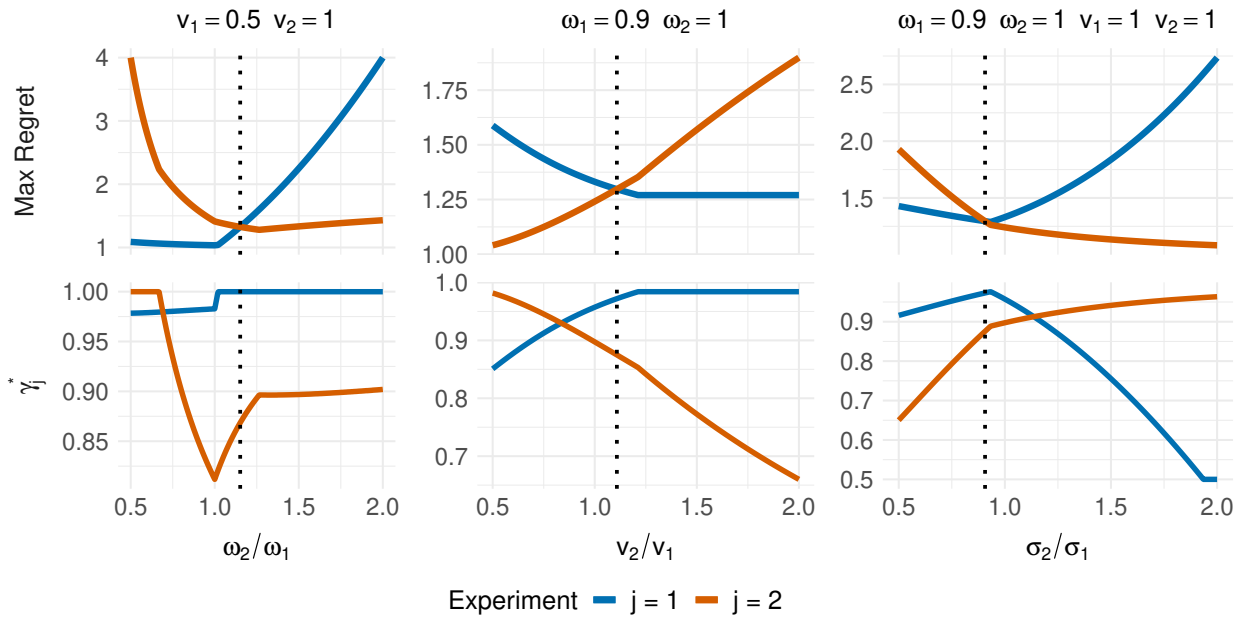

Figure 11: **Regret comparisons.** Top row: $\max\{\alpha(j,\gamma_j^\star)/\alpha^\star, \beta(j,\gamma_j^\star)/\beta^\star\}$ for $j \in \{1,2\}$ as the x–axis parameter varies (columns: ω_2, v_2, σ_2). The vertical dashed line marks indifference, where the two curves intersect; to its left/right, the optimal experiment is the one with lower max regret. Bottom row: optimal weights $\gamma_j^\star$ for $j = 1, 2$, which explain how sensitivity and precision interact to drive the design switch at $x^\star$. Columns vary, respectively, (a) ω_2 with $v_1 = 0.5$, $v_2 = 1$, $\sigma_1 = \sigma_2 = 1$, $\omega_1 = 1$; (b) v_2 with $\omega = (0.9, 1)$, $v_1 = 1$, $\sigma = (1, 1)$; (c) σ_2 with $\omega = (0.9, 1)$, $v = (1, 1)$, $\sigma_1 = 1$.

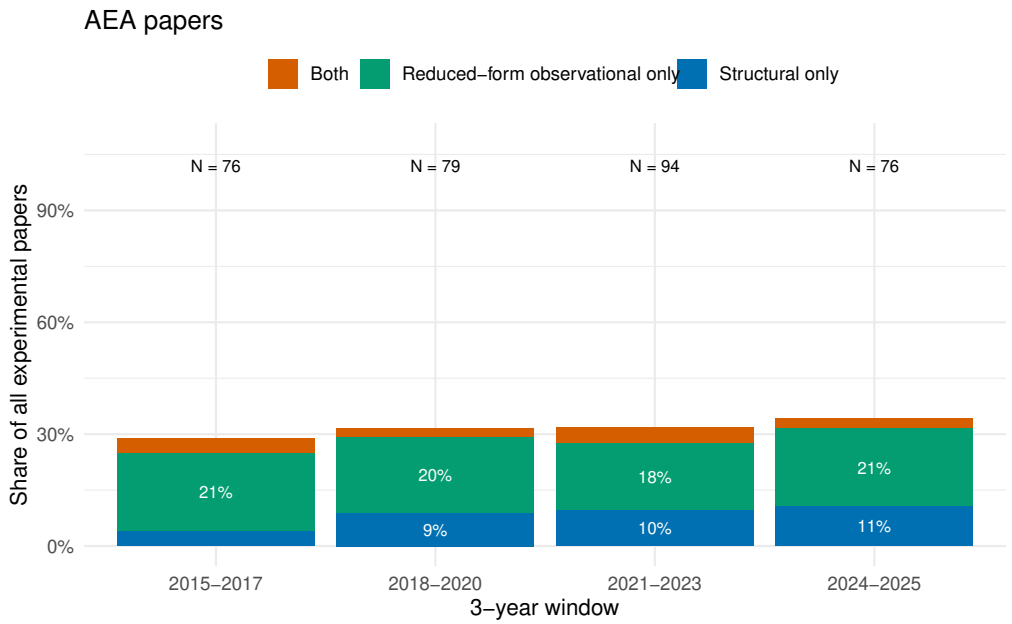

Figure 12: Share of experimental papers published in AEA journals also presenting experimental results in combination with observational estimates (either from a reduced form, or from a structural model or from both).

C Additional figures and tables

Table 5: Optimal shrinkage and experiment choice across different regimes in the presence of two parameters (Setting 2) The first three regimes (row) corresponds to the case where the variance ratio $\alpha/\alpha^\star$ does not uniformly dominate $\beta/\beta^\star$ for all values of γ_j and some $j \in \{1,2\}$ (which we view as leading cases). The other cases correspond to boundary solutions.

Case (condition)	**Optimal $\gamma_j^\star$**	**Optimal $j^\star$ solves**	**Solution $(j^\star, \gamma_j^\star)$**
Leading cases			
Case 1 no bias/ variance dominance $(\alpha(j,\gamma_j)/\alpha^\star - \beta(j,\gamma_j)/\beta^\star$ can flip sign)	$\gamma_j^\star$ in the interior for $j \in \{1,2\}$	$\arg\min_j \{\omega_{-j}^2\sigma_{-j}^2 + \omega_j^2[(1-\gamma_j^\star)^2\sigma_j^2 + (\gamma_j^\star)^2 v_j^2]\}$ $= \arg\min_j \{\lvert\omega_{-j}\rvert + \lvert 1-\gamma_j^\star\rvert\lvert\omega_j\rvert\}$	$j^\star$ minimizes bias and variance $\gamma_j^\star$ balances bias/variance
Case 2 $j=1$ is variance dominant $j=2$ has no bias/variance dominance	$\gamma_1^\star = \frac{\sigma_1^2}{v_1^2+\sigma_1^2}$ $\gamma_2^\star$ is interior	$\arg\min_j \{\omega_{-j}^2\sigma_{-j}^2 + \omega_j^2[(1-\gamma_j^\star)^2\sigma_j^2 + (\gamma_j^\star)^2 v_j^2]\}$	$j^\star, \gamma_1^\star$ minimizes variance and $\gamma_2^\star$ balances bias/variance
Case 3 $j=1$ is bias dominant $j=2$ has no bias/variance dominance	$\gamma_1^\star = 1$ $\gamma_2^\star$ is interior	$\arg\min_j \{\lvert\omega_{-j}\rvert + \lvert 1-\gamma_j^\star\rvert\lvert\omega_j\rvert\}$	$j^\star, \gamma_1^\star$ minimizes bias and $\gamma_2^\star$ balances bias/variance
Other (boundary) solutions			
Case 4 Both $j \in \{1,2\}$ are variance-dominant $(\alpha/\alpha^\star > \beta/\beta^\star$ for all γ and $j)$	$\gamma_j^\star = \frac{\sigma_j^2}{\sigma_j^2+v_j^2}$ for $j \in \{1,2\}$	$\arg\min_j \left\{\omega_{-j}^2\sigma_{-j}^2 + \omega_j^2\frac{\sigma_j^2 v_j^2}{\sigma_j^2+v_j^2}\right\}$	Minimizes variance
Case 5 Both $j \in \{1,2\}$ are bias-dominant $(\alpha/\alpha^\star < \beta/\beta^\star$ for all γ and $j)$	$\gamma_j^\star = 1$ for $j \in \{1,2\}$	$\arg\max_j \lvert\omega_j\rvert$	Minimizes bias
Case 6 bias dominant vs. variance dominant $(\alpha/\alpha^\star < \beta/\beta^\star$ for $j=1$ and vice versa for $j=2)$	$\gamma_1^\star = 1$ $\gamma_2^\star = \frac{\sigma_2^2}{\sigma_2^2+v_2^2}$	$\arg\min_j$ $\left\{\frac{\lvert\omega_2\rvert^2}{\lvert\omega_1\rvert^2}1\{j=1\} + (\omega_1^2\sigma_1^2 + \omega_2^2\frac{\sigma_2^2 v_2^2}{\sigma_2^2+v_2^2})1\{j=2\}\right\}$	Minimizes bias $\beta/\beta^\star$ of $j=1$ vs. variance $\alpha/\alpha^\star$ of $j=2$

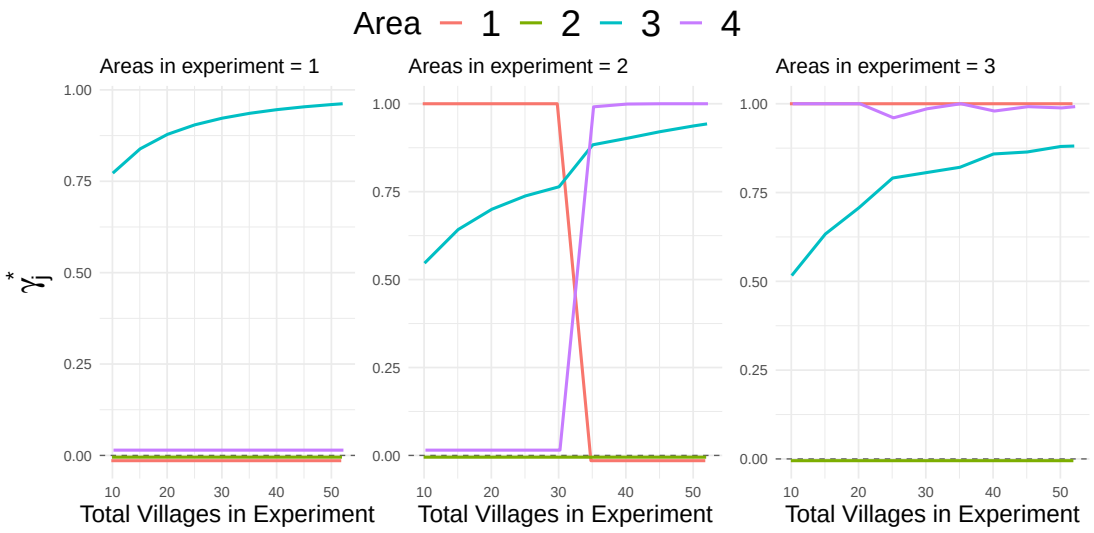

Figure 13: Optimal shrinkage $\gamma_j^\star$ by area as a function of total treated villages n_1 (from ten to fifty-two villages), under different constraints E. For $E < 4$, the solution is interior ($0 < \gamma_j^\star < 1$) and increases with n_1, approaching one as experimental noise falls.

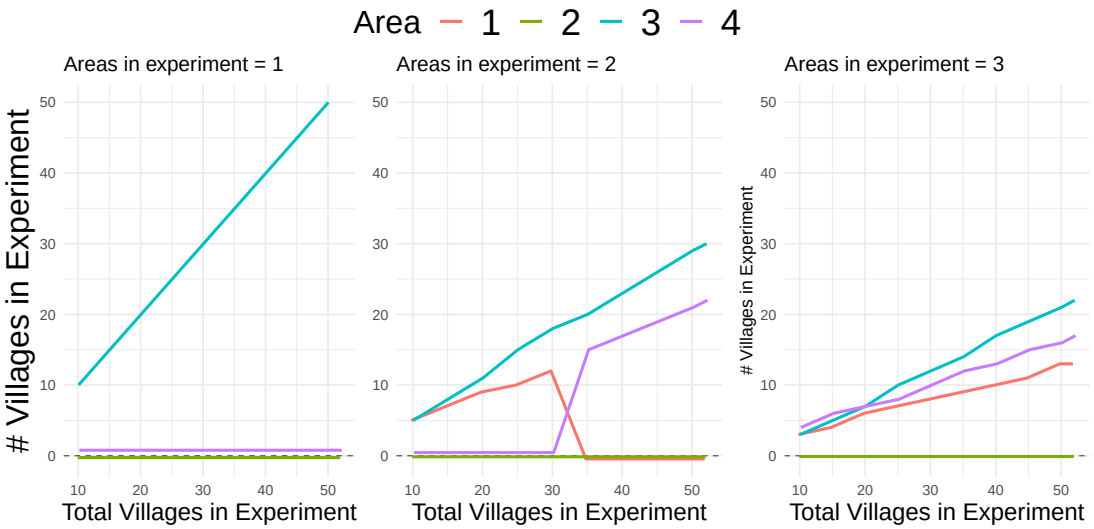

Figure 14: Area-level allocations as total treated villages n_1 varies (from ten to fifty-two villages), for each constraint on the number of eligible areas E. Lines show the number of villages assigned to each area.