# Highly Powered Analysis Plans *

Michael L. Anderson
UC Berkeley and NBER

Jeremy Magruder
UC Berkeley and NBER

April 1, 2022

**Abstract**

Formal analysis plans limit false discoveries by registering and multiplicity adjusting statistical tests. As each registered test reduces power on other tests, researchers prune hypotheses based on prior knowledge, often by combining related indicators into evenly-weighted indices. We propose two improvements to maximize learning within these types of analysis plans. First, we develop data-driven optimized indices that can yield more powerful tests than evenly-weighted indices. Second, we discuss organizing the logical structure of an analysis plan into a gated tree that directs type I error towards these high-powered tests. In simulations we show that researchers may prefer these "optimus gates" across a wide range of data-generating processes. We then assess our strategy using the community-driven development (CDD) application from Casey et al. (2012) and the Oregon Health Insurance Experiment from Finkelstein et al. (2012). We find substantial power gains in both applications, meaningfully changing the conclusions of Casey et al. (2012).

*Michael L. Anderson is Professor, Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720 (E-mail: mlanderson@berkeley.edu). Jeremy Magruder is Associate Professor, Department of Agricultural and Resource Economics, University of California, Berkeley, CA 94720 (E-mail: jmagruder@berkeley.edu). The authors gratefully acknowledge funding from the NSF under Award 1461491, "Improved Methodologies for Field Experiments: Maximizing Statistical Power While Promoting Replication," approved in September 2015. They thank Katherine Casey, Aprajit Mahajan, Ted Miguel, Sendhil Mullainathan, Ben Olken, and conference and seminar participants at University of Washington, Stanford, UC Berkeley, Notre Dame, University of Maryland, UT Austin, University of Wisconsin-Madison, and Cornell for insightful comments and suggestions and are grateful to Aluma Dembo and Elizabeth Ramirez for excellent research assistance. All mistakes are the authors'.

# 1 Introduction

A classic tradeoff in data analysis exists between estimating large numbers of parameters and generating results that do not reproduce in new samples. In computer science and machine learning this problem is known as "overfitting"; in biostatistics it manifests itself in "large-scale multiple testing." In the past decade it has become a critical issue in empirical microeconomics with the widespread use of field experiments.[1] Researchers designing field experiments often face high fixed costs in setting up the experiment and low marginal costs in adding additional survey outcomes. Increasing sample size is expensive, and the samples in many field experiments are too small to detect anything less than a large effect. Given these constraints and the focus on positive results in economics and other social sciences (Gerber and Malhotra 2008; Yong 2012), researchers face strong incentives to test for effects on many outcomes or subgroups and then emphasize the subset of significant results. Unfortunately this behavior maximizes the chances of "false discoveries" (type I errors) that do not replicate in new samples.

Economists have a range of tools available, both formal and informal, to limit false discoveries. Statistical methods that control the familywise error rate (FWER) or false discovery rate (FDR) formally test whether $p$-values are more extreme than would be expected under the null hypothesis based on the number of reported results. Credibly implementing these procedures, however, requires documenting the full set of conducted tests, usually through a preanalysis plan (PAP).[2] With a PAP, the researcher publicly documents the set of hypotheses that she intends to test prior to collecting the data, allowing formal control of type I error. Informal methods that limit false discoveries are also available; for example, registering a set of hypotheses through the AEA registry, or basing an analysis on a well-described (and perhaps registered) theory. Documenting the intent to pursue a particular direction of research demonstrates that the overall line of inquiry was not influenced by sample characteristics but allows researchers to respond to new ideas and information in the analysis, albeit at the cost of being unable to credibly control FWER at a specific value.

The tension between formal and informal control of false discoveries is driven by a tradeoff between statistical power and false discoveries. The fact that all hypotheses in a PAP must be

[1] It is also an issue in many observational studies, but it is difficult to establish when a researcher first had access to the data in an observational study. Establishing this timeline is critical to any method for limiting false discoveries.

[2] This method follows an approach used for decades in biostatistics (Simes 1986; Horton and Smith 1999) and appeared in economics at least as early as Neumark (2001). Casey et al. (2012) established best practices and popularized the use of PAPs among empirical microeconomics using a case of a Community-Driven Development (CDD) program in Sierra Leone. Both field experiments and PAPs have increased sharply in prevalence since 2010 (Currie et al. 2020).

anticipated constrains researchers and generates concerns about type II errors — failures to reject false hypotheses. Since each hypothesis included in a PAP that controls FWER increases test critical values for *other* hypotheses, power depends on the researcher correctly anticipating the set of hypotheses which are likely false and excluding those which are not. Individual researchers care about power, and the field as a whole suffers if novel discoveries are precluded or tests lack power to detect economically meaningful effects. Consolidating or otherwise reducing the number hypotheses can thus be attractive. Casey et al. (2012), for example, suggest combining outcome indicators into a range of unweighted summary indices to minimize the number of tests conducted. Olken (2015) recommends prespecifying a very small number of primary hypotheses, and foregoing formal FWER control over remaining hypotheses of interest. Banerjee et al. (2020) propose parallel streams for a "populated PAP", which reports the primary effects of an intervention, and a separate analysis in an academic paper, which foregoes control of false discoveries across the non-prespecified hypotheses. Such an approach balances the desire to demonstrate meaningful positive results on some indicators against the cost that novel findings may represent false discoveries.

This paper builds on Banerjee et al. (2020)'s insights to develop a key advantage of analysis plans: by formally controlling type I error we can generate statistical tests of correct size for *any* hypothesis. In doing so, we integrate and nest insights from classical econometrics, biostatistics, and machine learning (ML). We propose two tools to increase power on hypotheses of interest. First, we suggest an algorithm which maximizes power over the set of potential summary index hypotheses, allowing researchers to summarize many outcome variables in a single high-powered "optimus" test. As an index test, the optimus index retains the interpretation of an average treatment effect across multiple outcome indicators. Our approach strives to divorce anticipation of the most relevant indicators of interest from multiple inference control and instead uses the data to inform researchers as to which candidate index hypotheses are likely to represent high-powered tests. This approach leverages the analysis plan to avoid what would otherwise be a data-mining exercise resulting in tests with incorrect size. Second, we consider gatekeeping approaches to integrate the logical structure of an economic argument into the allocation of type I error. Gatekeeping approaches test hypotheses in serial rather than in parallel; in doing so they generate higher-powered tests on the first hypotheses examined at the potential cost of lower power on hypotheses tested subsequently. We demonstrate that combining gatekeeping with the optimized index generates a powerful test structure which yields substantial power gains over existing methods of controlling false discoveries. We expect this structure may be applicable in many economic contexts.

We consider two empirical applications. First, we reconsider the community-driven development (CDD) intervention studied by Casey et al. (2012). This application has several advantages:

it is the seminal application which popularized PAPs among microeconomists, and its PAP organizes indicators into families and suggests a clear logical structure for hypothesis tests. We adapt the pure PAP suggested by Casey et al. (2012) into a gatekeeping structure with optimized index tests. We conclude that the additional power generated by these tools would have led to important differences in the qualitative and quantitative understanding of the effects of the CDD program. Second, we consider the Oregon Health Insurance Experiment (OHIE), analyzed in Finkelstein et al. (2012). The analysis of OHIE was also prespecified. In contrast to Casey et al. (2012), however, the sample was large, effect sizes were more homogeneous, and the PAP-guided analysis yielded strong evidence in support of the effects of health insurance on healthcare utilization and some health outcomes. We thus treat the positive results presented in Finkelstein et al. (2012) as the true data generating process (DGP) and demonstrate that, in samples an order of magnitude smaller than the original sample, a gated optimus approach would have substantially higher statistical power to reject the null hypothesis than other available estimators.

The tools developed in this paper allow precise control of type I error and speak directly to two of the costs of formal analysis plans identified in Banerjee et al. (2020). First, statistical power may be greatly boosted by researchers using the optimus index over any other potential index that they could identify. Since the algorithm is prespecified, readers and reviewers need not worry about cherry-picking or the potential for false discoveries. Second, analysis plans based on these methods can be quite simple. To implement an optimus index test, one needs only know which indicators are grouped into which families of hypotheses. We suggest a simple gatekeeping structure which will be intuitive in many contexts: a first-stage gate that measures whether variables related to program implementation respond to treatment, a second-stage gate that tests one or more optimus indices in parallel, and a final stage that tests individual outcome indicators. Since the use of the optimus index generates high-power tests at each of the gates, researchers can concentrate error on each set of tests knowing that they have identified the highest power potential test to run. While the use of these tools does not replace the need for analysis and research beyond the registered analyses, the tools may greatly expand what can be learned through the rigorously-controlled PAP whenever power is not abundant.

While the optimus approach yields tests of the correct size, interpretation of effect sizes is a separate issue. We propose using a $K$-fold hold-out sample to form the optimus index, and demonstrate that the $K$-fold optimus regression coefficient is an unbiased estimator of a weighted average of treatment effects. The optimus weights towards outcomes with larger (true) treatment effects, however, so it does not represent the estimated effect size for the average outcome indicator (which could be reported separately). Nevertheless, as we demonstrate in our applications, the optimus

coefficient is often smaller in magnitude than the average coefficient on outcome indicators with positive results in a conventional PAP, because its higher power reduces the bias inherent in focusing on significant results (Andrews and Kasy, 2019).

The paper proceeds as follows. First, we set up a research environment in which researchers have access to a large number of outcome indicators (i.e. hypotheses) and are interested in testing for treatment effects on these outcomes. We then suggest that many of these indicators may be measurements of underlying latent variables, and discuss potential index tests deriving from these latent variable hypotheses. Next, we introduce the optimus index, and discuss gatekeeping approaches to error allocation. Section 3 describes numerical simulations, and Sections 4 and 5 discuss and present results for our applications. Section 6 concludes with recommendations.

# 2 Background

To structure the discussion, consider the case of a researcher who conducts a field experiment which assigns treatment, $T$, to a random fraction of the sample. For each participant $i$, she collects data on a set of $H$ outcomes, $\{Y_{i1}, Y_{i2}, ..., Y_{iH}\}$. These outcomes may be a mixture of individual variables and indices that aggregate multiple variables. They map to $H$ hypotheses, where the underlying relationship is

$$Y_{ih} = \beta_h T_i + \varepsilon_{ih} \tag{1}$$

The researcher wishes to test the null hypothesis $\mathbf{H}_h^0 : \beta_h = 0$ against the two-sided alternative $\mathbf{H}_h^A : \beta_h \neq 0$. Using the sample data, we can estimate the average treatment effect, $\hat{\beta}_h$, and an accompanying standard error, s.e.$(\hat{\beta}_h)$, that is an estimate of $\sigma_h$ (the standard deviation of $\hat{\beta}_h$). These are used to form a *t*-statistic under the null hypothesis, $\hat{t}_h = \frac{\hat{\beta}_h - 0}{\text{s.e.}(\hat{\beta}_h)}$. Using the *t*-distribution with $N-1$ degrees of freedom, the researcher can find a critical value of $\mathrm{t}_{\alpha/2}$.[3] If the estimated $\hat{t}_h$ falls above $\mathrm{t}_{\alpha/2}$ or below $-\mathrm{t}_{\alpha/2}$, we reject the null hypothesis $\mathbf{H}_h^0$ at the $\alpha$ significance level. As scientific convention, we take $\alpha = 0.05$.

The set $\mathcal{H} = \{1, ..., H\}$ enumerates all candidate outcome variables $Y_h$, where $h \in \mathcal{H}$ is associated with a hypothesis as described above. In most field experiments the implementation of the treatment is expensive, but measuring an additional outcome has low marginal cost. Often $H$ is therefore large.

We denote the benchmark objective function as the *Simple Rejection Problem*. In the Simple Rejection Problem, the researcher maximizes the expected sum of statistically significant treatment

[3] Let $t \sim t_{N-1}(0,1)$ be distributed according to the centered *t*-distribution with $N-1$ degrees of freedom and standard deviation of 1. The probability of $t$ falling anywhere above the critical value $\mathrm{t}_{\alpha/2}$ or below $-\mathrm{t}_{\alpha/2}$ is $\alpha$.

effects. This objective function accords with one of the definitions of power that Romano et al. (2010) propose (p. 95), and we use it throughout the paper. The researcher forms expectations about rejections according to a prior belief $F_h$ over $\{\beta_h, \sigma_h\}$ and selects a subset of hypotheses to test, $\mathcal{H}' \subseteq \mathcal{H}$, that solves

$$\max_{\mathcal{H}' \in 2^{\mathcal{H}}} \mathbb{E}\left[\sum_{h \in \mathcal{H}'} \mathbb{I}\{|\hat{t}_h| > \mathrm{t}_{\alpha/2}\}\right] = \max_{\mathcal{H}' \in 2^{\mathcal{H}}} \sum_{h \in \mathcal{H}'} \mathbb{P}_{F_h}\left(|\hat{t}_h| > \mathrm{t}_{\alpha/2}\right) \tag{2}$$

There is no constraint in the maximization problem above, so the maximizing subset, $\mathcal{H}^*$, is the subset of hypotheses with a positive probability of rejection. Since even true hypotheses reject at rate $\alpha$, the maximizing subset is $\mathcal{H}^* = \mathcal{H}$, and the researcher tests for effects on every possible outcome. This solution naturally opens the door to false discoveries, and limiting these false discoveries is a critical issue in most empirical disciplines (Sterling 1959).

## 2.1 False Discovery Problem

The fundamental problem with testing every hypothesis in $\mathcal{H}$ is that in any hypothesis test there is a chance that the sample statistic falls in the rejection region, even if the null hypothesis is true. This false discovery problem leads to costly but ultimately futile future research, as well as potentially dangerous policy. More broadly, it erodes the trust that the public has in the results that researchers find. Thus it is important to minimize the rejection of true hypotheses, or the type I error rate.[4]

Returning to the researcher's decision in Equation (2), in the worst-case scenario all the null hypotheses in $\mathcal{H}$ are true. Even though the study contains no false hypotheses, it still rejects $\alpha \cdot |\mathcal{H}|$ of the hypotheses in expectation. As an example, suppose 100 hypotheses are tested at a significance level of 0.05. Even if all 100 null hypotheses are true, we expect the study to (incorrectly) reject five of the null hypotheses, generating five significant findings.

To address this issue, multiplicity adjustments work to control the overall type I error rate of the study. This error rate is either the probability that the study makes at least one incorrect rejection — the familywise error rate — or the expected proportion of rejections that are incorrect — the false discovery rate. The simplest adjustment is the Bonferroni correction, which controls FWER. With the Bonferroni correction, we divide $\alpha$ by the number of hypotheses tested, in this case, $|\mathcal{H}'|$.[5]

[4]This paper is not the first to discuss the false discovery problem in the context of randomized experiments in economics or the general social sciences. For example, see Anderson (2008), Anderson and Magruder (2017), and Fafchamps and Labonne (2017) for related discussions of these issues and techniques for controlling the type I error rate.

[5]More sophisticated adjustments exist that minimize the power reduction associated with additional tests. Nevertheless, it is inherent in the control of FWER, or the probability of making any type I error (i.e. false rejection),

The researcher's problem becomes[6]

$$\max_{\mathcal{H}' \in 2^{\mathcal{H}}} \mathbb{E}\left[\sum_{h \in \mathcal{H}'} \mathbb{I}\{|\hat{t}_h| > \mathrm{t}_{\alpha/2|\mathcal{H}'|}\}\right] = \max_{\mathcal{H}' \in 2^{\mathcal{H}}} \sum_{h \in \mathcal{H}'} \mathbb{P}_{F_h}\left(|\hat{t}_h| > \mathrm{t}_{\alpha/2|\mathcal{H}'|}\right) \tag{3}$$

where $\mathrm{t}_{\alpha/2|\mathcal{H}'|}$ is the critical value above which a standard *t*-statistic has probability $\frac{\alpha}{2|\mathcal{H}'|}$ of falling.

The critical value $\mathrm{t}_{\alpha/2|\mathcal{H}'|}$ increases with $|\mathcal{H}'|$; for example, $\mathrm{t}_{\alpha/2|\mathcal{H}'|} = 3.49$ if $|\mathcal{H}'| = 100$. In this example, a hypothesis that would reject with 80% probability prior to multiplicity adjustment — a common benchmark in study design — would reject with only 24% probability after multiplicity adjustment. The more hypotheses the researcher tests, the higher the critical value becomes, and the lower the probability of rejecting a given hypothesis becomes.

A straightforward response to this tension is to reduce the dimensionality of $\mathcal{H}'$, and a frequently utilized tool to do so is to aggregate related indicators into a small number of index hypotheses (e.g. Kling et al. (2007)). Ideally, this preserves the economic result identified by the test while paying a double dividend for power: it reduces the number of hypotheses tested *and* generates indices with smaller standard deviations than their underlying components.

At the same time, whether the researcher aggregates indicators or not, the validity of the multiplicity adjustment requires honest disclosure of $\mathcal{H}'$, which creates an incentive problem for researchers. A researcher motivated to increase rejections could test every hypothesis in $\mathcal{H}$ but report a subset, $\mathcal{H}_r$, that contains only hypotheses with large *t*-statistics. In many cases $|\mathcal{H}_r| << |\mathcal{H}|$, and the multiplicity adjustment for each test becomes much less severe.[7] Thus, multiplicity adjustments are only effective when researchers can credibly communicate the number of hypotheses they have tested.

Historically, biostatistics has taken a strong interest in controlling false discoveries. This interest arises from the large financial incentives and potential welfare impacts related to false discoveries in clinical trials and the massive number of hypotheses tested in many genomics studies. It has thus become standard practice in the medical literature that clinical trials should register analysis plans prior to enrolling patients (De Angelis et al. 2004). More recently, empirical microeconomics has begun to adopt this model for field experiments in the form of preanalysis plans.

---

that adding more tests requires more stringent adjustment of *p*-values. Otherwise, the probability of making at least one error rises. The only case in which FWER would not rise would be the case in which the new test is perfectly correlated with one or more of the existing tests. In this case the new test does not represent new information.

[6] Here $\mathbb{I}\{\cdot\}$ is the indicator function, equal to 1 if the condition $\{\cdot\}$ is true, and equal to 0 otherwise.

[7] When there exist many candidate index hypotheses, the problem is arguably greater: researchers motivated by rejections can fine-tune the selection of indicators into the index with the best in-sample performance.

## 2.2 Preanalysis Plans

One way to credibly communicate the number of hypotheses to be tested is to file a preanalysis plan. A PAP describes in detail the analyses that a researcher intends to perform, including the construction of any index hypotheses. An effective PAP requires that the researcher upload it to a public site, such as the AEA RCT Registry, prior to collecting her data. With a publicly registered PAP, the researcher "ties her hands" with respect to the analysis, thus preventing "cherry picking" of results or "$p$-hacking." Formally, readers can be confident that the reported set of tested hypotheses, $\mathcal{H}_r$, represents the true set of tested hypotheses, $\mathcal{H}'$. In what follows, we define the *exhaustive PAP* to indicate the PAP which prespecifies every hypothesis in $\mathcal{H}$, that is the PAP where $\mathcal{H}_r = \mathcal{H}' = \mathcal{H}$.

In addition to specifying the hypotheses to be tested, an effective PAP must specify some form of multiplicity adjustment for statistical tests (assuming it tests more than one hypothesis). Without any multiplicity adjustment, the researcher's optimal strategy is to include as many hypotheses as possible, even those that may be very unlikely to reject or of little interest, since the option value of including any given hypothesis test in the PAP is weakly positive. The constraints on the PAP thus become the researcher's creativity and value of time.

Multiplicity adjustments formalize the implicit tradeoff that motivates PAPs to begin with. Each additional test has option value in that it may reject and be of interest, but it also carries an explicit cost in that it reduces the power of other included tests. These adjustments thus impose discipline on the researcher's hypothesis selection process.

## 2.3 Aggregate Indices in PAPs

In many contexts, a number of indicator variables may correspond to the same latent economic or conceptual hypothesis. In such cases, we may be able to partition the hypothesis set $\mathcal{H}$ into $G$ groups, such that $\mathcal{H}_g$ contains the hypotheses belonging to group $g$, and treatment effects are anticipated to be weakly monotonic within a group.[8] Then we can represent the data generating

[8]This weak monotonicity assumption, expressed in Equation (4) as $\gamma_{hg} \in [0, c)$, assists with interpretation of the estimated treatment effect, but the test we propose maintains the correct size even if weak monotonicity fails. Nevertheless, higher power estimators that do not impose weak monotonicity may be possible if the indicators are not weakly monotonic.

process as

$$\begin{aligned} y^*_{ig} &= \delta_g T_i + \varepsilon^*_{ig} \\ y_{ihg} &= \gamma_{hg} y^*_{ig} + u_{ihg} \; \forall \; h \in \mathcal{H}_g \\ \gamma_{hg} &\in [0, c) \end{aligned} \tag{4}$$

Now $\beta_{hg} = \delta_g \gamma_{hg}$. In these cases, the researcher may be primarily interested in rejecting the hypothesis $H_g : \delta_g = 0$; then, rejecting any convex combination of the indicators in $\mathcal{H}_g$ suffices to reject $H_g$. This insight motivates reducing the problem of dimensionality in the hypothesis space through the construction of an aggregate index hypothesis, which are often unweighted averages of outcomes (Kling et al., 2007). That is, researchers test $\beta_{\bar{y}} = 0$ by estimating the regression

$$\frac{1}{|\mathcal{H}_g|} \sum_{h \in \mathcal{H}_g} y_{ihg} = \beta_{\bar{y}} T_i + \nu_{ig} \tag{5}$$

for hypotheses in group $g$. More generally, they may test a weighted version of the index

$$\mathbf{w}'_g \mathbf{y}_{ig} = \beta_{\mathbf{w}_g} T_i + \nu_{i\mathbf{w}_g} \tag{6}$$

where $\mathbf{y}_{ig}$ is a $|\mathcal{H}_g| \times 1$ column vector of outcomes in group $g$, $\mathbf{w}_g$ is a $|\mathcal{H}_g| \times 1$ column vector of weights summing to one (and otherwise unrestricted), and $\beta_{\mathbf{w}_g}$ and $\nu_{i\mathbf{w}_g}$ are scalars. They may use generalized least squares (GLS) weights to increase power (O'Brien, 1984; Anderson, 2008); then, $\mathbf{w}'_g = (\mathbf{1}'\Sigma_g^{-1}\mathbf{1})^{-1}(\mathbf{1}'\Sigma_g^{-1})$, where $\mathbf{1}$ is a column vector of ones and $\Sigma_g$ is the covariance matrix for $\mathbf{y}_{ig}$.

In practice, many authors follow Kling et al. (2007) in using unweighted mean indices across all outcomes in a group $g$. In what follows, we define the index containing an unweighted average of all standardized outcomes in group $g$ as the *KLK index* for that group. This is a natural choice: if hypotheses are homogeneous with respect to $\beta_{hg}$ and $\Sigma_g$ (i.e. all outcomes have the same standardized treatment effects and are equally correlated with each other), the KLK index maximizes statistical power (see Corollary 1.2 in Appendix A2.1). On the other hand, if there is dispersion in the distribution of $\beta_{hg}$ or the elements of $\Sigma_g$, then an index based off the hypotheses with larger treatment effects or lower covariances can be more powerful than the KLK index. Of course, designing such an index would require researchers to correctly anticipate the vector $\beta_{hg}$ and matrix $\Sigma_g$, which is often unrealistic.

Even if the latent index representation is not literally correct, mean indices have several appealing qualities. They allow researchers to infer that the regression coefficient is the effect on a (weighted) average of outcomes in a well-specified index. If the member hypotheses in the index

have similar treatment effects, then estimating Equation (5) leads to a higher-powered test than testing any individual hypothesis in isolation (see Corollary 1.1 in Appendix A2.1). Furthermore, it allows the $p$-values to be corrected only for the number of groups $G$, not the number of indicator hypotheses $H$. Nevertheless, mean indices are not costless. In a PAP, they induce the possibility of type II error if researcher priors are inaccurate. In particular, if the researcher does not correctly anticipate the full vector of indicator effects and covariance matrix of those effects, then they may design indices which increase the likelihood of type II error (i.e. have lower power than testing individual indicators, even after multiplicity adjustment).

## 2.4 Optimized Aggregate Indices

The use of an analysis plan opens the opportunity to identify and test other index hypotheses. Specifically, an analysis plan can define an algorithm allowing the data to suggest a high-powered index hypothesis. Suppose that a researcher receives utility from rejecting $H_g : \delta_g = 0$, the family-level hypothesis in Equation (4), but no additional utility from rejecting individual indicator-level hypotheses $H_{hg} : \beta_{hg} = 0$. Then the researcher will wish to maximize power across possible indices composed of elements of $\mathcal{H}_g$, the index "donor pool". We define the "optimus index", which maximizes power across potential indices for group $g$:

$$\max_{\mathbf{w}_g} \mathbb{P}_{F_{\mathcal{H}_g}} \left( |\hat{t}_{\bar{y}_{\mathbf{w}_g}}| > \mathrm{t}_c \right). \tag{7}$$

Deriving the optimus index requires knowledge of the DGP, which we denote $F_{\mathcal{H}_g}$. Specifically it requires the coefficients $\beta_{hg}$ and the covariance matrix $\Sigma_g$ for all $h \in \mathcal{H}_g$. Let $\beta_{\mathbf{g}}$ be a $|\mathcal{H}_g| \times 1$ column vector containing $\beta_{hg}$ for all $h \in \mathcal{H}_g$. In Appendix A2.1 we derive the following proposition for known $\beta_{\mathbf{g}}$ and $\Sigma_g$:

**Proposition 1.** Consider an index $\bar{y}_{i\mathbf{w}_g} = \mathbf{w}_g' \mathbf{y}_{ig}$. Let $\hat{\beta}_{\mathbf{w}_g}$ be the regression coefficient from estimating Equation (6) and let $\sigma_{\hat{\beta}_{\mathbf{w}_g}} = \sqrt{\mathrm{V}(\hat{\beta}_{\mathbf{w}_g})}$. A one-sided test of $\beta_{\mathbf{w}_g} = 0$ based on $\hat{\beta}_{\mathbf{w}_g} / \sigma_{\hat{\beta}_{\mathbf{w}_g}}$ with critical value $\Phi^{-1}(1-\alpha)$ has power $\Phi(\frac{\beta_{\mathbf{g}}' \mathbf{w}_g}{\sqrt{\mathbf{w}_g' \Sigma_g \mathbf{w}_g}} + \Phi^{-1}(\alpha))$.

Absent detailed priors researchers generally default to the KLK index, which contains an unweighted average of all indicators in the donor set $\mathcal{H}_g$. Access to sample estimates, however, can generate a higher powered test. A natural approach for deriving the optimus index is to replace $\beta_{\mathbf{g}}$ and $\Sigma_g$ with sample estimates $\hat{\beta}_{\mathbf{g}}$ and $\hat{\Sigma}_g$. Doing so, however, results in an overestimate of the test's power (and the treatment effect size), since the procedure heavily weights outcomes with the largest *t*-statistics, which would likely experience mean reversion in a hold-out sample. In essence, there is an overfitting problem; specifically, the estimates $\hat{\beta}_{hg}$ and $\hat{\Sigma}_g$ are overdispersed relative to

the true $\beta_{hg}$ and $\Sigma_g$ (the largest $\hat{\beta}_{hg}$ is likely large both because the true $\beta_{hg}$ is large and because it experienced a sampling error shock of the same sign as $\beta_{hg}$). In Appendix A2.2 we demonstrate formally that utilizing the full sample to estimate $\hat{\beta}_{hg}$ and $\hat{\Sigma}_g$ forming the optimus using these estimates results in a biased estimator:

**Proposition 2.** Let $\omega_g = \text{argmax}_{\mathbf{w}_g} \Phi(\frac{\hat{\beta}'_{\mathbf{g}}\mathbf{w}_g}{\sqrt{\mathbf{w}_g\widehat{\Sigma}_g\mathbf{w}_g}} + \Phi^{-1}(\alpha))$, where $\hat{\beta}_{\mathbf{g}}$ and $\widehat{\Sigma}_g$ are sample estimates of $\beta_{\mathbf{g}}$ and $\Sigma_g$. Let $\beta_{\omega_g} = E[\beta_{\mathbf{g}}{}'\omega_g \mid \omega_g]$. Consider an index $\bar{y}_{i\omega_g} = \omega'_g\mathbf{y}_{ig}$. A regression of $\bar{y}_{i\omega_g}$ on $T_i$ yields a biased estimate of $\beta_{\omega_g}$.

To address this bias we incorporate several machine-learning techniques when deriving the optimus index. First, we utilize sample splitting. With sample splitting, researchers can estimate $\hat{\beta}_{hg}$ and $\hat{\Sigma}_g$ in a training sample and then apply the derived optimus index in a test sample. For the simulations and applications we incorporate 5-fold sample splitting (Hastie et al. 2009, p. 242). For each fold, we estimate the optimus index using the data that omits that fold, and then apply the estimated optimus index weights to the omitted fold. Aggregating these indices across folds generates an optimus test that can be implemented on the full sample. As different folds of the data may feature different constructions of the optimus index, rejecting the optimus test using a $K$-fold approach implies that there is a mean treatment effect on a subset of variables in group *G*, where the weights of the specific component indicators may be summarized across the full sample.

In Appendix A2.2 we formally demonstrate that estimating Equation (6) using the $K$-fold version of the optimus test produces an unbiased estimator of the expected weighted average of the elements of $\beta_{\mathbf{g}}$ , with weights determined by the $K$-fold procedure:

**Proposition 3.** Randomly assign $N$ observations to $K$ folds. For each fold $k$, compute weights $\omega_{-k,g} = \text{argmax}_{\mathbf{w}_{-k,g}} \Phi(\frac{\hat{\beta}'_{-\mathbf{k},\mathbf{g}}\mathbf{w}_{-k,g}}{\sqrt{\mathbf{w}_{-k,g}\widehat{\Sigma}_{-k,g}\mathbf{w}_{-k,g}}} + \Phi^{-1}(\alpha))$, where $\hat{\beta}_{-\mathbf{k},\mathbf{g}}$ and $\widehat{\Sigma}_{-k,g}$ are estimates of $\beta_{\mathbf{g}}$ and $\Sigma_g$ using all observations not in fold $k$. Let $\tilde{\mathbf{T}}$ be a demeaned $N \times 1$ vector of treatment assignments and $\tilde{\mathbf{Y}}_g$ be a $N \times 1$ vector of weighted outcomes, with element $i$ equal to $\omega'_{-k,g}\mathbf{y}_{ig}$. The $K$-fold optimus estimator $(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\tilde{\mathbf{Y}}_g$ is unbiased for $E[\beta_{\mathbf{g}}{}'\omega_{-k,g}]$.

Proposition 3 states that the optimus $K$-fold procedure is unbiased for a weighted average of treatment coefficients $\beta_{\mathbf{g}}$, with weights equal to the expected optimus $K$-fold weights. Furthermore, the average weights across folds, $\bar{\omega}_g = \frac{1}{K}\sum_k \omega_{-k,g}$, represent an unbiased estimate of the expected weights due to the random assignment of folds. Thus in our applications we report the average optimus weight (across folds) that each outcome receives.

Nevertheless, the *t*-statistic for the $K$-fold optimus estimator is not distributed *t*, because each fold is ultimately used both to form the optimus weights and to estimate the treatment effect (see

Corollary 3.1 in Appendix A2.2). Therefore, test statistics from this approach should be tested against critical values generated by randomization inference (that is, randomly permuting treatment across the full sample and implementing the procedure on these random treatment permutations many times).

While the $K$-fold optimus is unbiased, the overdispersion in estimates of $\beta_{\mathbf{g}}$ may reduce the finite sample efficiency of the estimator. To counteract the overdispersion that tends to arise in estimates of $\beta_{\mathbf{g}}$, we modify the objective function in Proposition 1 to include a penalty for indices that concentrate weight on a smaller number of indicators. Specifically, we evaluate

$$\max_{\mathbf{w}_g} \Phi\left(\frac{\hat{\beta}'_{\mathbf{g}}\mathbf{w}_g}{\sqrt{\mathbf{w}'_g\widehat{\Sigma}_g\mathbf{w}_g}} + \Phi^{-1}(\alpha)\right) - \lambda HHI_{\mathbf{w}_g} \tag{8}$$

where $HHI_{\mathbf{w}_g}$ represents the Herfindhal-Hirschmann index (HHI) for weights $\mathbf{w}_g$ (i.e. $\sum_{h\in\mathcal{H}_g} w_{hg}^2$). This penalty index, which by construction must lie on the unit interval, encourages the optimus test to be a well-defined index hypothesis which presents average treatment effects across a range of variables rather than, for example, selecting the single indicator with the most significant *t*-statistic. In the simulations, we experiment with a range of values of $\lambda$ to determine which penalty weight generates the highest power index across different DGPs. In the applications we apply the preferred penalty weight from the simulations.[9]

Next, the off-diagonal elements of the estimated covariance matrix, $\hat{\Sigma}_g$, may be particularly overdispersed, and small or negative off-diagonal entries can have substantial effects on the indices' predicted power. To address this overdispersion we derive an Empirical Bayes shrinkage estimator for $\hat{\Sigma}_g$ and use it to shrink the off-diagonal elements of $\hat{\Sigma}_g$ in our applications (see Appendix A3).

Finally, it is common in the existing literature to assume that effects on individual indicators, appropriately transformed, are weakly monotonic. This weak monotonicity assumption leverages the underlying latent index model of Equation (4). Formally, it implies that the index loadings, $\gamma_{hg}$, are weakly positive (or weakly negative). Enforcing weak monotonicity is also appealing because a mixture of positive and negative weights confounds the directional interpretation of the weighted index, regardless of the underlying DGP.[10] Thus researchers may wish to restrict the optimus weights to be weakly monotonic, as we do in our simulations and applications. The only

[9] In an actual application one could also tune the penalty weight using cross-validation. Doing so in our applications, however, would be computationally prohibitive, in part because we draw multiple samples to explore performance in different scenarios.

[10] The possibility of a mixture of positive and negative weights is a key reason why researchers often avoid the GLS-weighted index (Pocock et al., 1987; Dallow et al., 2008).

downside to this restiction is that if the true loadings are not weakly monotonic, then an index that mixes positive and negative weights could be more powerful, though less interpretable.

These approaches together leverage an analysis plan to identify an index test which is highly powered, which controls type 1 error, and which is an unbiased estimator of a well-defined weighted average of treatment effects across variables belonging to a given family. The cost of doing so is that the researcher is not guaranteed a test of the unweighted average across indicators comprising the KLK index; instead, the data determines which outcome variables are most strongly associated with treatment in a way that the researcher need not anticipate *ex ante*. The benefit of doing so is a reduction in type 2 error. We explore the extent of these benefits by simulation in Section 3 and in applications in Sections 4 and 5. The costs of using the optimus test depend on the difference in inherent interest between the KLK index and the set of well-defined weighted average indices which could exist on outcome variables in family $g$. In investigations where the researcher selects the KLK index due to a well-defined theory that suggests homogeneous treatment effects across outcomes in family $g$, these costs could be significant. On the other hand, in investigations where the researcher anticipates heterogeneous treatment effects but selects a KLK index to maximize statistical power because they have uninformed priors, they may be quite small.

The optimus test has clear analogues in other tests that have been implemented or proposed in the literature. For example, while the optimus test maximizes the expected *t*-statistic of the index in the confirmation sample ($E[\hat{\beta}/\hat{\sigma}]$), O'Brien's GLS weights minimize the standard error ($E[\hat{\sigma}]$). As such, if treatment effects are uniform among the hypotheses in a group, the two should converge to the same weighted index. Similarly, while the optimus test focuses on maximizing power to detect mean treatment effects on a subset of indicators in the data, the machine learning based test in Ludwig et al. (2019) (hereafter LMS) flexibly tests the sharp null hypothesis of any treatment effects across the marginal and joint distributions of outcomes.[11] The finite sample performance of the two estimators, as well as the DGP (i.e. whether the primary treatment effects lie on the mean effects of treatment or on the joint distribution of outcomes), determine which of these two procedures has greater power. As in most econometric applications, imposing additional structure can improve precision if that structure is consistent with the DGP. In our applications we test the optimus index alongside the LMS procedure.

[11] In the Ludwig et al. (2019) case, the researcher inverts the problem and uses split sample and ML methods to predict treatment using outcome variables.

## 2.5 Optimizing Error Allocations through Gatekeeping

A second contributor to type II error that challenges PAPs in economics is the misallocation of type I error. Tests which limit false discoveries by controlling type I error often treat hypothesis tests concurrently and uniformly. This strategy may lead to allocating type I error to some hypotheses which are of researcher interest only conditional on other rejections; if these latter rejections fail to materialize, then type I error goes wasted.

Gatekeeping strategies define the propagation of type I error across hypotheses. The key insight is that when a test rejects, its type I error can be recycled to another test in a prespecified manner. Formally, a gatekeeping strategy controls FWER at level $\alpha$ across sequential families of hypotheses $F_1, ..., F_M$. Each family represents a "gate" that must be passed. In a serial gatekeeping strategy, hypotheses in family $F_j$ are tested iff all hypotheses in family $F_{j-1}$ are rejected using *p*-values that are multiplicity adjusted within family $j-1$. In a parallel gatekeeping strategy, hypotheses in family $F_j$ are tested iff at least one hypothesis in family $F_{j-1}$ is rejected using *p*-values that are multiplicity adjusted within family $j-1$ (Dmitrienko and Tamhane 2007). Tree-structured gatekeeping strategies may be most relevant to field experiment practitioners (Dmitrienko et al. 2007; Bretz et al. 2011), as they allow researchers to precisely specify how type I error flows between hypotheses.

As a simple example, consider a field experiment with imperfect compliance and a relatively small sample. A reasonable tree-structured gatekeeping strategy in this context could specify three families: $F_1, F_2, F_3$. $F_1$ contains the first-stage *t*-statistic, $F_2$ contains one or more aggregate index tests, and $F_3$ contains the outcomes comprising the aggregate indices. The researcher first tests $F_1$ with no multiplicity adjustment. If $F_1$ rejects — i.e. there was a first-stage effect — she then tests for any aggregate effect on outcomes via the test(s) in $F_2$. Failure to reject $F_1$ and $F_2$ precludes testing of individual outcomes, but in this context it is unlikely the researcher could generate compelling findings absent a first-stage or overall effect. If an aggregate index in $F_2$ rejects, the individual indicators in $F_3$ comprising that index could be tested in parallel.[12]

[12]If the outcomes of the field experiment can be partitioned into $G$ groups, then a reasonable tree-structured gatekeeping strategy in this context could specify four families: $F_1, ..., F_4$. $F_1$ contains the first-stage *t*-statistic, $F_2$ an optimus index test across all outcomes in the study, $F_3$ optimus or KLK index tests for each group $g$ of the $G$ groups, and $F_4$ the indicators comprising those indices. The researcher first tests $F_1$ with no multiplicity adjustment. If $F_1$ rejects — i.e. there was a first-stage effect — she then tests for any aggregate effect on outcomes using the optimus index in $F_2$ — a high-powered test suited to her small sample. If $F_2$ rejects she then tests for effects on the $G$ groups in $F_3$, multiplicity adjusting for $G$ tests. Finally, any group that rejects would have its component indicators tested in $F_4$, with $\alpha/G$ type I error to allocate across all indicators in that group. Failure to reject a family at any stage precludes testing of subsequent families.

As the example makes clear, using gatekeeping changes the nature of misallocation error. Without the use of gates, researchers spread type I error across many hypotheses, running the risk of reducing power on false hypotheses by including true ones. Using a gate allows researchers to concentrate type I error on a high-powered test, but that advantage comes with a cost: failing to reject the gate prevents formal testing of the indicators comprising the index test within the gate. Gatekeeping methods are therefore likely to be most useful when there exist tests that have a high probability of rejecting under the alternate hypothesis (e.g. an optimus index) and when the value of some rejections increases conditional on other rejections.

For researchers with latent index hypotheses like Equation (4), the power advantages of aggregate index hypotheses render them as natural gates. For a researcher with a prespecified plan and incomplete knowledge of the underlying DGP, writing a PAP that uses KLK indices as gates may be sensible. For two reasons, however, we anticipate that using optimus tests as gates will have significant advantages for many researchers. First, since the optimus gate maximizes statistical power among tests of mean treatment effects, using an optimus test as a gate both maximizes the chance of producing statistical evidence for a mean treatment effect across a group of variables and minimizes the risk of failing to pass the gate, which would preclude tests of component indicators. The simulations and applications below explore the potential power differences between KLK index and optimus gates. Second, some research designs may be complex, with many potential sets of hypothesis families. In this case, anticipating a network and path for the propagation of type I error across families and hypotheses quickly becomes intractable (Olken, 2015; Banerjee et al., 2020). By selecting indicators within a family which have the strongest relationship to treatment, optimus gates control type I error over indices that are most related to treatment. In many cases, the optimus test may help simplify analysis design by identifying an index of variables with strong treatment effects instead of requiring the researcher to anticipate this set.

# 3 Analysis Plan Simulations

When combining our test strategies with FWER control procedures more sophisticated than the Bonferroni correction, it is infeasible to analytically calculate power. We thus turn to Monte Carlo simulations to evaluate the performance of different strategies across a wide range of potential data generating processes. For the optimus test, the simulations also give us insight into reasonable values for the HHI penalty weight in Equation (8).

## 3.1 Simulation Environment

We perform a series of Monte Carlo simulations that establish the power of our strategies relative to KLK indices or an exhaustive PAP under a variety of scenarios. In this context we use "power" to refer to the probability that a single test rejects or, when considering multiple tests, the expected number of rejections. Power depends on some parameters that the researcher has direct control over (number of tests, use of an aggregate index or gatekeeping strategy), some that she has limited control over (sample size), and others that she has no control over (share of hypotheses that are false, effect sizes, and inter-test correlation structure).

Effect size and sample size are fundamental to statistical power. These two factors interact to generate the sampling distribution of the test statistic, which determines power. The question of what *t*-statistics a researcher might expect to find thus informs her expected power. To limit the parameter space of interest we conducted a literature review of field experiments with the goal of determining the empirical distribution of published *t*-statistics (described in Appendix A1). This literature review concluded that the median *t*-statistic in published field experiments was 2.6. We thus simulated DGPs in which the expected *t*-statistic for a false hypothesis, $E[t_h \mid \beta_h \neq 0]$, ranged from 1.5 to 4.0.

To assess the performance of optimus indices and gatekeeping strategies across a range of contexts, we set up the following simulation environment. First, there are $H$ outcomes with $H$ corresponding hypotheses. Of these $H$ hypotheses, $H_1$ are false, and the remainder true. False hypotheses have a normalized mean "effect size" of $\mu_t = E[t_h \mid \beta_h \neq 0]$, where the data-generating process draws a coefficient $\beta_h$ using the degenerate distribution (homogeneous treatment effects) or from a gamma distribution with shape parameter $2\mu_t$ and scale parameter $\mu_t/2$ (heterogeneous treatment effects). True hypotheses have $\beta_h = 0$. A fraction $r$ of outcomes are correlated with correlation coefficient $\rho$, generating correlated tests.

Let the $H \times 1$ column vector $\boldsymbol{\beta}$ represent the $H$ coefficients. To test for robustness in a broad range of environments we vary total hypotheses ($H$), the number of false hypotheses ($H_1$), average effect size ($\mu_t$), inter-outcome correlations ($\rho$), the share of outcomes that are correlated ($r$), and the coefficient DGP (degenerate or gamma distributions) across simulations (see Table 1).

To simulate a $K$-fold optimus index, we draw $K = 5$ column vectors of coefficients, each dimension $H \times 1$, centered at $\boldsymbol{\beta}$. Each element in each vector has variance $K$, such that the average of $\hat{\beta}_{hk}$ across all $K$ vectors — i.e. the "full-sample coefficient" — has unit variance. The full-sample coefficients are thus distributed standard normal around $\boldsymbol{\beta}$ and can be treated as *t*-statistics. We generate two aggregate indices from the $H$ outcomes. One is a KLK index that includes all $H$ outcomes, equally weighted. The second is an optimus index that solves Equation

(7). Due to computational constraints, in the simulations we only consider *unweighted* optimus indices, i.e. those in which the non-zero weights are identical.[13] The KLK index is based off of full-sample coefficients, $\hat{\boldsymbol{\beta}}$. The optimus index is derived $K$ times using the $K$ folds. For each fold $k$, the optimus index is derived using $\hat{\boldsymbol{\beta}}_{-k}$, i.e. coefficients estimated while omitting fold $k$, and then applied to fold $k$. We average the resulting $K$ optimus indices across the $K$ folds. We consider optimus objective functions (Equation (8)) that apply values of $\lambda = 0$, 0.01, 0.1, 0.5, 1, 2, and 4 for the HHI penalty weight.

For gatekeeping purposes we apply either the KLK index or the optimus index as an initial gate. For large $H$ this structure simulates a scenario in which the index tests for any effect study-wide and serves as a gate for the entire study; for small $H$ it simulates a scenario in which the index tests for an effect on a subgroup and serves as a gate for that subgroup. If the gate rejects, we test all the coefficients in $\boldsymbol{\beta}$. We also simulate exhaustive PAPs by testing all the coefficients in $\boldsymbol{\beta}$ without any indices, and we simulate "parallel plans" in which we test an index in parallel with all the coefficients in $\boldsymbol{\beta}$ (i.e. we simultaneously conduct $H + 1$ tests). We correct for multiple hypothesis testing with a Romano-Wolf (RW) algorithm that controls FWER.[14] To ensure the correct test size for our $K$-fold optimus index, we simulate the null distribution when setting $\boldsymbol{\beta} = \mathbf{0}$ and reject based on that distribution.

## 3.2 Simulation Results

Table 1 presents the different parameter values used in the simulations. We simulate power — i.e. the expected number of rejections — for 2,600 combinations of parameter values in total. In the discussion we also focus on "more empirically relevant" parameter values, which include combinations for which $\mu_t \leq 3$ and $H_1/H \leq 0.5$ (i.e. studies with moderate power), based on surveys we conducted of field experiments and PAPs (see Appendix A1). Results for the optimus index depend in part on the value of the HHI penalty weight, $\lambda$. In general average power across different parameter values did not vary strongly with $\lambda$, but overall the optimus appeared to perform best with $\lambda = 0.5$. We thus report results using $\lambda = 0.5$ for the simulations and applications.

We first consider the scenario in which a researcher wishes to perform an aggregate index or

[13] Accordingly, we benchmark the optimus against an unweighted KLK index as opposed to, for example, a GLS-weighted index.

[14] To run these simulations we generate positively correlated test statistics. Most FWER control procedures that incorporate dependence between test statistics, such as the free step-down resampling method or the step-wise method in Romano and Wolf (2005), rely on resampling to determine the correlation structure. Resampling is computationally infeasible in our simulations, so we instead developed a rejection-region FWER control method based off the results in Romano and Wolf (2005) that leverages the known correlation structure of our DGP.

omnibus test. Table 2 characterizes the tradeoffs between using an optimus index or a KLK index. Column (1) reports average power over all parameter combinations, while Columns (2) through (5) report average power over parameter combinations that are more empirically relevant. Columns (3) and (4), in particular, focus on small families ($H \leq 20$) and large families ($H \geq 50$) respectively. Column (3) thus simulates a scenario in which a researcher tests an index corresponding to a subset of hypotheses (e.g. educational outcomes in a conditional cash transfer experiment that measures effects on educational, health, and financial outcomes), while Column (4) simulates a scenario in which a researcher tests for any treatment effect across all outcomes.

Table 2 reports average power of a $K$-fold optimus index relative to a KLK index. Row 1 summarizes the case in which the researcher tests the index in isolation; thus there is no multiplicity adjustment. In this case, the optimus test is between 1.8 and 3.5 times more powerful than the KLK index on average. Row 2 summarizes the case in which the researcher tests the index in parallel with an exhaustive PAP, multiplicity adjusting all tests. The optimus index averages between 3.6 and 15.9 times the power of the KLK index, with orders of magnitude gains in power when there are many hypotheses. In summary, as power becomes more scarce (with increasingly heavier multiplicity adjustments), the advantage of the optimus index becomes more stark.

Row 3 summarizes the average size (number of indicator variables) of an optimus index for each set of parameter combinations. Across all parameter combinations the optimus contains an average of 17.1 variables (Column (1)). Among smaller families, the optimus averages 4.9 variables (Column (3)), and among larger families it averages 18.5 variables (Column (4)). The results demonstrate that the optimus, while smaller than the KLK index, still tends to capture effects averaged across at least 5 to 20 variables.

The results in Table 2 suggest that researchers should generally prefer the optimus index unless they get many times more utility from rejecting the KLK index. The only case in which the KLK index averages more than half the power of the optimus index is the first row entry — i.e. an index test in isolation — in Column (1). In this scenario, however, it is unlikely that the researcher wishes to test only one hypothesis (the index) across the entire study. More likely, the index serves as a gate that, if passed, allows the researcher to test other hypotheses. When the index serves as a gate, then power becomes more important since rejecting the gate opens the door to further tests. In that context a KLK index is typically unattractive (see Appendix Table A3).

Researchers may test the optimus index in a serial (i.e. as a gate) or parallel (i.e. in conjunction with many other hypotheses) fashion. Table 3 analyzes the relative power of a $K$-fold optimus-gated plan versus a plan that tests the $K$-fold optimus index in parallel. The optimus-gated plan first tests an optimus index for evidence of any treatment effect; if the index rejects, it then tests

prespecified individual outcomes. The optimus-parallel plan tests the same hypotheses, including the optimus index, but conducts all tests simultaneously (i.e. there is no gate). A value of 1.00 in a cell implies the two strategies have identical power. Each row corresponds to a different weight that the researcher assigns the index test, with a weight of 1.0 implying that rejecting the index is of equivalent value to rejecting a single outcome.

Regardless of weight, the optimus-gated plan is higher power on average than the optimus-parallel plan, with power advantages increasing as the optimus index becomes more intrinsically interesting. The intuition is that when few hypotheses are false, the power of the optimus index is much higher than the power of the typical indicator variable, and it is beneficial to concentrate type I error on the index (via gating). Alternatively, when many hypotheses are false, the optimus index is highly powered, and there is little downside to using it as a gate. Moreover, the strong preference to use the index as a gate when the optimus test is more interesting (i.e. its weight is higher) is clear: if the optimus is just another indicator to test, then increasing the odds of rejecting the optimus (at the cost of being unable to reject subsequent indicators if the optimus fails to reject) can yield only modest gains. On the other hand, if rejecting the index test is more valuable than rejecting a single outcome, then there are substantial benefits to concentrating power on the index test first and, if it rejects, recycling the type I error to test individual outcomes.

To illustrate the distribution of relative power, Figure 1 plots histograms of the relative power of an optimus-gated PAP against a KLK index-gated PAP. Figures 1a and 1b correspond to Columns (1) and (2) of Table 2 respectively.[15] The optimus-gated PAP dominates the KLK index-gated PAP in most cases across all parameter combinations and virtually all cases across more empirically relevant parameter combinations. Figures 1c and 1d plot distributions for small and large families respectively (Columns (3) and (4) of Table 2). The mean power advantage is higher for large families, but in both cases the optimus-gated plan dominates the KLK index-gated plan for virtually all of the plotted parameter combinations. Figures 1e and 1f reproduce 1c and 1d but apply double weight to rejecting the KLK index, relative to the optimus index (a weight of 4 versus 2). Even with double weight on the KLK index, the optimus-gated plan continues to dominate the KLK index-gated plan in most scenarios. Overall, the figure suggests that researchers will prefer an optimus-gated plan over a KLK index-gated plan in most cases, and almost universally prefer it when working with large families.

For researchers deciding whether to test any index at all, Figure 2 plots histograms of the relative power of optimus-parallel and optimus-gated plans, with index weights of 1 and 3, against

[15] In this figure we apply an index weight of 2 to the optimus index (i.e. rejecting the optimus is twice as interesting as rejecting a single indicator) to ensure that a gatekeeping plan is preferred over a parallel test plan. The KLK index also receives a weight of 2, except in the last two panels, where it receives a weight of 4.

an exhaustive PAP with no index tests. Figures 2a and 2b plot distributions for an optimus-parallel plan in which the index receives weights 1 and 3 respectively (with a weight of 1 again implying that rejecting the index is of equivalent value to rejecting a single outcome). The mean power advantage is higher when the optimus weight is higher, but in both cases the optimus-parallel plan dominates the exhaustive PAP for all parameter combinations. Figures 2c and 2d reproduce the first two panels but switch to an optimus-gated plan (which is generally preferred over the parallel plan). The advantage of the optimus-index plans becomes even more decisive.[16]

In summary, the simulations suggest that adding an optimus test in parallel to a PAP uniformly dominates ignoring the optimus test for researchers who place any nontrivial value on rejecting the optimus test. Furthermore, our results indicate that adopting an optimus-gated plan will be superior to the parallel plan for most researchers, with particularly large gains for those who find the (optimus) index test to be of intrinsic interest or who anticipate relatively low power in their analysis plan. This straightforward conclusion belies the multidimensional nature of these research strategies, with each dimension potentially interacting with the others. In principle researchers may choose no index, an optimus index, or a KLK index, and a gatekeeping test strategy or a parallel test strategy. Appendix A4 examines the relative power of the different combinations of research strategies. It finds that, on average, plans with indices outperform plans without indices, and plans with optimus indices outperform plans with (unweighted) KLK indices. Thus, in most cases, the best strategy for a researcher is to include an optimus index test, typically as a gate.

# 4 Application: GoBiFo Revisited

Casey et al. (2012) document the impacts of GoBiFo, a community-driven development (CDD) intervention in Sierra Leone. To control false discoveries in a survey collecting hundreds of outcomes, they developed a preanalysis plan comprising of 12 KLK index hypotheses. We summarize the Casey et al. (2012) discussion of the institutional features of GoBiFo here before noting several features of the evaluation that make it an appealing choice of an application.

CDD programs are an important outlet for international donor funding, and GoBiFo had a variety of features common to CDD-type programs in the developing world. First, it provided block grants, training, and business start-up capital based on community proposals with a goal of

[16]Plans that include indices have a natural advantage over those that do not because they are testing an extra hypothesis. In some cases, however, an outcome may reject both as part of an index and by itself. Appendix Figure A1 reproduces Figure 2 but includes a double-rejection adjustment, described in Appendix A4, that ensures that researchers do not receive extra utility from rejecting the same hypothesis twice. While the distribution of power ratios shifts left, the plans which include indices dominate an exhaustive PAP in virtually all cases.

enhancing public-goods access. These grants were substantial relative to local living standards: financial outlays were $4,667 per village, or about $100 per household. To receive these grants, village development committees (VDCs) were required to submit a development proposal to the ward development committee (WDC), the next higher level of government bureaucracy, for review, endorsement, and transmission to the relevant District Council for approval. 43% of grants were used for local public goods (such as community centers, sports fields, primary school repairs and sanitation); 40% applied to agriculture and livestock or fishing management (such as seed multiplication, communal farming, or goat herding); and the remaining 17% went towards skills training and small business development initiatives. Casey et al. (2012) describe these facets of the GoBiFo intervention as the "hardware" of the intervention.

On top of block grants to create new public goods, GoBiFo had several features meant to build democratic institutions, which may be particularly relevant in the traditional authority context of Sierra Leone. In particular, GoBiFo both established VDCs, which would play a role in coordinating local governance, and instituted participation requirements for historically marginalized groups, such as women and youth. These participation requirements included, for example, that VDC bank accounts included at least one female signatory and that public works proposals document evidence of the inclusiveness of women and youth in the proposal generation requirements. Inclusiveness and democratization were monitored by GoBiFo staff at substantial cost — monitoring and facilitating this institution building cost about as much as the actual development grants given out. Casey et al. (2012) describe this facet of GoBiFo as the "software" effects of the CDD program.

Casey et al. (2012) introduce a PAP with twelve KLK index hypotheses, listed in Table 4. The PAP also specifies *t*-statistics and FWER-adjusted $p$-values, reported in the paper. These 12 index hypotheses are split into two groups. The first three hypotheses relate to the "hardware" of public-goods provision in the village, and in all three cases Casey et al. find strong evidence that the "hardware" of public-goods provision changed. Examining the underlying variation, these hypotheses confirm that GoBiFo was successfully implemented and led to an outlay of funds and investment in public goods. The remaining nine hypotheses relate to the "software" of the program, examining a range of outcomes, including participation in collective action, trust of leaders, participation in local governance, and reductions in crime and conflict in the community. Casey et al. (2012) find no evidence that GoBiFo affected any of these outcomes, at least after adjusting *p*-values for the number of hypotheses tested (twelve). Ultimately they conclude that the program was implemented as planned and led to some expenditures and a change in the public-goods environment, but that there is no evidence that it changed the social institutions governing these

villages.

The evaluation of GoBiFo makes for a natural test application of our methods for several reasons. First, as one of the seminal papers introducing preanalysis plans to economists, it represents a carefully thought out and well-regarded PAP that has become a template for PAPs in the literature. Second, the results are mixed. On the one hand, the study had more than adequate statistical power to detect effects on the hardware hypotheses. However, the null effects on the software hypotheses may reflect either a lack of impact on these institutions or a lack of statistical power to detect these effects, given the number of prespecified hypotheses and the indicators selected to join each index hypothesis. As such, there may be an opportunity to learn more about the impacts of this program if procedures with greater statistical power can be leveraged. Finally, the PAP slots naturally into a gatekeeping environment. Casey et al. clearly delineate two meta-hypotheses: that GoBiFo was implemented successfully and influenced the "hardware" of public goods provision, and that GoBiFo influenced the institutional "software" that underlies public goods provision. In our framework, this suggests a natural gate structure to the hypotheses which we develop below.

We begin by replicating results from Casey et al. (2012). The twelve hypotheses in Casey et al. (2012) are each average treatment effects across the whole sample, estimated by comparing endline treatment and control outcomes. Thus, the primary results in the PAP and the initial presentation come from estimating

$$y_v = \beta T_v + \gamma X_v + \varepsilon_v$$

where $X_v$ are covariates used for stratification and $T_v$ is an indicator for treatment status. In all cases, the outcome variable $y_v$ refers to a KLK index hypothesis, constructed by normalizing and summing indicator variables which relate to a particular hypothesis about the intervention (following O'Brien (1984) and Kling et al. (2007)).

## 4.1 Alternate Indices and Gates in GoBiFo

The hypotheses presented in GoBiFo are heterogeneous in the number and types of outcome variables which enter each index. Inputs to index hypotheses include outcomes verifiable through administrative data, objective and subjective survey data questions, and behaviors elicited through "supervised community activities" — for example, direct observation of how community members stored and shared a tarpaulin gifted by the survey team. These different data sources lead to heterogeneous index variables, with as few as seven and as many as 47 variables averaged in the construction of each hypothesis's index. In this context, it seems plausible that treatment effects would be heterogeneous within indices, and that there exist alternate specifications for these indices which might have also yielded similar levels of intrinsic interest.

To generate a gated optimus index approach, we propose that a logical gating meta-hypothesis is whether there are any impacts on the (aggregate) set of hardware hypotheses that serve as a "first stage" test of GoBiFo. If GoBiFo had no impacts on any of the indicators in the hardware section — i.e. the program was not implemented and did not deliver public goods — readers would have good cause for skepticism about any positive statistical results that follow. If GoBiFo was implemented successfully, so that some hardware indicators responded to treatment, then it is plausible that the program might have impacts on institutional software as well. Researcher priors over which hardware elements are critical to generate software change, however, may not be detailed; as such, conducting an optimus-index test for "were there hardware effects?" may serve as a high-powered gate.

The second meta-hypothesis refers to software effects. If GoBiFo had hardware effects, it is natural to test next whether there were any software effects. As such, conducting an optimus test over all the variables which comprise the nine software hypotheses represents a high-powered test of this meta-hypothesis.[17]

Finally, if we conclude that there were both hardware and software effects, we may wonder which of the 12 finer hardware and software hypotheses were impacted by GoBiFo. Once again, if researchers do not have strong preferences over the weights assigned to indicator variables, they can construct an optimus-style index of the variables belonging to each hypothesis and anticipate greater statistical power, relative to regressions using KLK indices, based on the results in Section 3.2.[18]

Figure 3 summarizes the gating structure we use for GoBiFo. For each of the index hypotheses — hardware effects, software effects, and the 12 prespecified hypotheses — we implement the optimus approach using 5-fold CV. Since it is impossible for us to "preregister" the fold assignments, we generate results using many different fold assignments and record the distribution of $p$-values across the different fold assignments. Specifically, we assign five folds at random 200 times, stratifying each draw of five folds on treatment. In each of these 200 iterations, we compute $p$-values by comparing actual test statistics to those generated by the same procedure when randomly permuting treatment under the null hypothesis 80 times. When examining multiple comparisons (in the

[17] In practice, the GoBiFo PAP prespecified a number of variables which contribute to multiple different hypotheses. In some cases, a variable appears in both hardware and software hypotheses. For our implementation, we eliminate all indicator variables that appear in any hardware hypothesis from the set of candidate software variables.

[18] If some of the hypotheses reject, the individual indicator hypotheses can then be tested in sequence if the larger index gate hypotheses are rejected. To preserve the correct size of tests in this case, however, one needs to avoid using sharpened $p$-values, because sharpened $p$-values recycle type I error within a parallel set of tests until it is exhausted (e.g. one could not apply Romano-Wolf $p$-values). For brevity, we do not present those results here.

analysis of the 12 hypotheses) we compute *p*-values using the Romano and Wolf (2005) algorithm. Following Chernozukhov et al. (2018) and Romano and DiCiccio (2019), we conservatively reject a hypothesis only if the median *p*-value of that hypothesis across the 200 5-fold assignments is less than $\alpha/2$. For optimus indices, we apply a HHI penalty weight of $\lambda = 0.5$ in the objective function (based on results from Section 3).

We benchmark these results against three alternatives. First, we consider the results presented in Casey et al. (2012) based on their original PAP, which tests all hypotheses in parallel. Second, we construct a gatekeeping version of the original PAP, which uses KLK indices for hardware and software outcomes. The hardware index includes all variables from Hypotheses 1–3, and the software index includes all variables from Hypotheses 4–12 (omitting any variables also classified as hardware). Third, we apply the Ludwig et al. (2019) omnibus procedure to the same gates. To implement LMS we use an ensemble that combines a random forest with an elastic net to predict treatment using the hardware dependent variables as a first gate; if we pass that gate then we do the same for software variables. As Ludwig et al. (2019) also requires a sample split, we follow the same procedure as for the optimus tests, estimating the LMS procedure on 200 sets of five folds and computing *p*-values via permutation of the treatment indicator.

## 4.2 GoBiFo Results

In Table 5 we consider whether we reject the null of no hardware effects using the original PAP, KLK index, optimus gate, and LMS approaches. All four approaches reject the hardware gate. We therefore conclude that using any of these approaches would have (correctly) concluded that GoBiFo had an effect on hardware outcomes, though the specific interpretations of each rejection differ (as discussed in Section 2.4).

Since each approach rejects the hardware gate, we then consider whether GoBiFo had impacts on software variables. Recall that the original PAP for GoBiFo failed to reject the null hypothesis that GoBiFo had no effect on software outcomes. Combining all software variables into a single KLK index does not change this conclusion, as shown in Column (1) of Table 5. Column (3) similarly demonstrates that the LMS omnibus test does not reject the null hypothesis of no relationship between software variables and treatment; the median *p*-value is 0.34. In contrast, the optimus approach, reported in Column (2), rejects the null hypothesis of no software effects. Using the 5-fold optimus, the median *p*-value is 0.00, so that we reject the null hypothesis that there was no relationship between GoBiFo and software variables at the conventional 5% significance level. On average, the optimus test produces an index that is a weighted average of 22.8 indicator variables, 19 of which receive an average weight greater than 2.5%, and none of which receive an average

weight above 10%.[19] Appendix Table A5 reports the variables appearing most frequently in the software optimus and their associated average weights.

Casey et al. (2012) note that one software variable — whether there is a community farm — may be miscategorized as software, as it may have been directly built by the CDD grant. The third row repeats the software optimus gate but excludes this variable; it still rejects the null based on an average of 22 variables.

In summary, when applying an optimus test we conclude that GoBiFo affected some software outcomes. Comparing point estimates between the KLK index (Column (1)) and the optimus (Column (2)) illustrates why we see such a large difference. Column (1) indicates that GoBiFo is associated with a 0.03 standard deviation average increase across 144 distinct software indicators; Column (2) indicates that across the 22.8 components of the optimus test, the weighted average effect is 0.15 standard deviations. Notably, the gap between the two estimates is of similar magnitude to what we would expect if the treatment effects on all indicators excluded by the optimus were zero.

Using the gated approach, only the optimus test has sufficient power to pass the software gate. After passing the gate we can then test which of the underlying hypotheses contribute to this rejection. We construct 12 optimus indices for the 12 underlying hypotheses and test them in parallel, computing *p*-values using the Romano and Wolf (2005) algorithm. As presented in Table 4, we find that the optimus gate approach rejects each of the three individual hardware hypotheses with large weighted average effect sizes. Among the software hypotheses, only Hypothesis 6 ("GoBiFo changes local systems of authority, including the role and perception of traditional leaders (chiefs)") approaches marginal significance; adjusting for the 12 tested hypotheses, the median Romano-Wolf *p*-value of 0.0625 would not quite reject at the 10% level based on the conservative Chernozukhov et al. (2018) bound.[20]

Examining the indicator variables that receive the greatest weight in the software optimus index (Appendix Table A5) yields insights as to why we find limited evidence to reject individual software hypotheses. First, the five most heavily weighted variables each correspond to different underlying software hypotheses. Summing across all indicator components of the optimus, the

[19] We count an indicator as appearing in the index if it receives at least as much weight as it would in a KLK index; in this case that corresponds to $0.007 = \frac{1}{144}$.

[20] The optimus test for both Hypothesis 6 and Hypothesis 4, "Participation in GoBiFo increases collective action and contributions to public good" have median naive *p*-values below 0.025, so that either of these might reject in an analysis plan with fewer prespecified hypotheses. Interestingly, the optimus test for H4 yields a greater than 70% weight on the community farm variable, lending support to the hypothesis that that variable is misclassified hardware (included indicators and weights for each hypothesis are presented in Appendix Table A5).

single hypothesis receiving the greatest aggregate weight is the marginally-significant Hypothesis 6; weights on indicators comprising Hypothesis 6 sum to 27% of the total weight.[21] This pattern suggests that the overall software effects were spread across the identified software hypotheses rather than concentrated within one of them. Rather than belonging to a specific hypothesis proposed in the PAP, what stands out in the variables selected for heavy weights is that they tend to focus on more objective measurements rather than subjective indicators, such as those which examine beliefs and attitudes. In addition to the presence of a community farm (Hypothesis 4), heavily weighted indicators include whether the respondent has been in a recent physical fight (H11); whether minutes were taken at a recent village council meeting (H5); whether they are a member of a women's group (H8); whether newly elected chiefs were young (H6); and whether the respondents can accurately name the year of the next general election (H9). This trend indicates that the optimus is identifying treatment effects that exist across an index of relatively objective measurements, perhaps because of measurement error in subjective assessments.

We conclude that GoBiFo had a meaningful effect on a subset of the software outcomes, that that effect was distributed across prespecified hypotheses about software, and that of the considered approaches only the gated optimus test had the statistical power to detect it.

# 5 Application: The Oregon Health Insurance Experiment

Finkelstein et al. (2012) examine the effects of a 2008 health insurance lottery in Oregon on health care utilization, financial well-being, and self-reported health outcomes. In 2008 Oregon identified sufficient financial support to expand access to "OHP Standard" — a Medicaid-expansion offering — to an additional 10,000 potential beneficiaries. To identify these beneficiaries, the state proposed to allocate the plan by lottery among the 89,824 applicants who registered from eligible households. The state selected 35,169 potential beneficiaries by lottery, of whom 30% successfully enrolled in Medicaid. In comparing the randomly-selected beneficiaries to those who were not selected, Finkelstein et al. (2012) combine rich administrative and survey data to provide causal evidence on the effects of health insurance on health care utilization and financial and health outcomes.

Finkelstein et al. (2012) follow a prespecified analysis plan. The PAP estimates the equation

$$y_{ihj} = \beta_0 + \beta_1 lottery_h + X_{ih}\beta_2 + \varepsilon_{ihj} \tag{9}$$

[21] Two of the heavily-weighted indicators are components of both Hypothesis 6 and another hypothesis (in one case Hypothesis 5, and in another Hypothesis 12); the repeated use of the same indicators across hypotheses indicates some of the challenges in partitioning hypotheses in institutional analysis.

where $y_{ihj}$ represents outcome $j$ for individual $i$ in household $h$, $lottery_h$ indicates that household $h$ was a lottery winner, and $X_{ih}$ are covariates that determine the probability of winning the health-insurance lottery (household size and survey-round fixed effects).[22] Finkelstein et al. (2012) report several key findings. First, access to health insurance boosted health-service utilization. Using both administrative and survey data, lottery winners had more inpatient stays and outpatient visits and were more likely to receive prescription drugs. They also engaged in more preventative care, undergoing more cholesterol tests, high blood sugar tests, mammograms, and Pap smears. Consistent with their utilization, they reported better access to health care: they were more likely to have a usual clinic, a personal doctor, and to report receiving all needed medical care and prescription drugs. Finally, access to health care had a positive impact on perceived health: lottery winners reported being in better health, both physically and mentally.[23]

The OHIE study, with a sample size in the tens of thousands, had more than adequate statistical power — many *t*-statistics for individual indicators are on the order of 5 to 10. We thus treat OHIE as an opportunity to test the performance of our techniques in a context in which we know the "true" DGP. Specifically, we sample a small fraction of the OHIE data and compare the power of an optimus-gated analysis plan to plans gated by a KLK index or the LMS omnibus test and to an exhaustive PAP that tests all outcomes in parallel (with no gate). We then verify that the conclusions are consistent with the true DGP.

Figure 4 summarizes the structure of the OHIE analysis plan. As with GBF, a logical gating meta-hypothesis is the existence of a first-stage effect — absent any effect on insurance status, it is implausible that the lottery affected other outcomes. Assuming there is a first-stage effect, we can then test whether any of the outcomes, including those related to utilization, financial strain, and self-reported health, were affected. Finally, conditional on insurance having some effect, we can test individual indicators to determine which were affected. We implement the optimus approach using the same 5-fold CV algorithm described in Section 4.1.[24]

[22] Specifications using administrative data also include covariates to improve precision. The administrative data, however, are not publicly available.

[23] Interestingly, Finkelstein et al. (2012) point out that a number of patterns suggest that this improvement in health is not likely directly attributable to health service utilization, as these changes appear in survey data well before any differences in health service utilization emerge.

[24] There are two subtle divergences from the GBF-analysis procedure. First, since we are not generating substantive results for OHIE, we assume that the researcher could preregister the CV folds for the analysis, removing the need to generate many sets of folds and apply the conservative Chernozukhov et al. (2018) bound for the median *p*-value. Second, since OHIE treatment is only random conditional on covariates, we stratify the null treatment permutations on covariates as well (i.e. we ensure that the average treatment probability in each covariate cell, after permutation, matches the original treatment probability in that cell). This stratification is critical for generating tests of the correct

For each targeted sample size we draw 100 random samples, executing the analysis 100 times. We consider samples that are 8%, 10%, 12%, and 15% of the original sample. At each sample size we ask what power an exhaustive PAP, a gated KLK index approach, a gated optimus approach, or the LMS omnibus test would have. We focus on the survey data outcomes, as nearly all of the administrative data outcomes are not publicly available. For simplicity we estimate intention-to-treat (ITT) effects.[25] We populate the family of outcomes using all measures listed in the original PAP that could plausibly respond to treatment. The outcomes fall into three broad categories: care-seeking behaviors, including outcomes related to health care utilization, preventative health care, and health care access; self-reported health outcomes; and financial outcomes. To ensure that our conclusions comparing different approaches are not specific to the pooling of all outcomes into a single family, we also explore the power of the optimus index and KLK index when applied separately to each of the three subfamilies.[26] Appendix Table A6 lists the indicators used across all three subfamilies, alongside measurements of the ITT effects and FWER-adjusted $p$-values from the full (100%) sample.[27] The table reveals that at least nine (of the 44) individual indicators reject in the 100% sample, even when controlling FWER across all 44 tests, with at least one rejection in each of the three subfamilies.

## 5.1 OHIE Results

Table 6 reports average rejection rates for different families (rows) across different analysis plans (columns). We focus on the 10% sample because it represents a scenario in which power approaches, but does not reach, the 80% rule-of-thumb target. Appendix Table A8 presents analogous results for the 8%, 12%, and 15% samples. The first-stage test for an effect on Medicaid coverage is identical across all analysis plans. The lottery strongly increased Medicaid coverage in

size.

[25] Instrumental variables estimates are approximately 3.5 times the ITT estimates; as documented in Section 5.1 the first-stage estimation error is trivial.

[26] We note that these families are somewhat different from the indices reported in Finkelstein et al. (2012), who report standardized effects at the sub-table level, often consisting of only two or three indicators. We adopt the larger families to focus on highlighting statistical properties at much smaller subsamples, where aggregating over more indicators would be attractive for statistical power. Nevertheless, we also report results for several small, homogeneous table-level families defined in the OHIE PAP that formed the basis of Finkelstein et al. (2012).

[27] Our goal with the OHIE data is to compare the performance of different analytic strategies rather than to establish novel substantive results. Thus we limit the 100% sample to individuals with complete data for the outcome indicators we identify ($N = 8,141$), rather than imputing data for missing outcomes. We take the 100% sample as the "true" effects in the sense that they represent the estimands for estimates based on random subsamples of the data. These estimates may not be unbiased for the true ITT effects, however, if OHIE outcome data are not missing at random.

the original study ($F = 1,930$), and the first-stage rejects with 100% frequency in every sample.

Since all approaches reject the first stage, we then consider whether each approach finds that OHIE impacted one or more of the 44 plausible outcomes. Column (1) reports the power to reject this hypothesis for each approach. The optimus approach (correctly) concludes that OHIE affected outcomes 71% of the time. This represents 22% higher power than the (unweighted) KLK index, which rejects the null 58% of the time. The LMS omnibus test rejects 38% of the time, suggesting that its additional flexibility is not helpful in this context. Finally, an exhaustive PAP that tests 44 indicators in parallel rejects one or more indicators 58% of the time, with a median rejection of one indicator (conditional on rejecting anything).

Column (2) of Table 6 reports average effect sizes where relevant. As expected, the average effect size for the optimus index, 0.09 standard deviations, is larger than the KLK index average index effect size of 0.04 standard deviations. The gap between the two effect sizes is less pronounced than in GoBiFo, reflecting the smaller family size and more modest effect heterogeneity in OHIE. The gap between the two estimates is also smaller than what we would expect if the treatment effects on all indicators excluded by the optimus were zero, suggesting that in this context the optimus chooses indicators with larger effect sizes rather than all indicators with nonzero effects.

On average, the optimus procedure constructs an index that is a weighted average of 9.0 indicator variables (Column (3)), six of which receive an average weight greater than 4%, and none of which receive an average weight above 20%.[28] Appendix Table A7 reports the variables appearing most frequently in the optimus and their associated average weights. Outcomes that stand out include indicators for (not) paying any out-of-pocket medical costs in the past six months, reporting the usual place of care is a clinic, having any primary care visits, and getting all needed medical care in the past six months.

Column (4) of Table 6 reports the "true" optimus and KLK index effect sizes, based on the 100% sample. To compute these estimates we apply the average optimus index weights to generate a weighted index in the 100% OHIE sample and then estimate Equation (9) in the 100% sample using this weighted index as the outcome.[29] We do the same for the KLK index index but use identical weights for each indicator variable. The estimates in Columns (4) confirm that both procedures estimate the correct effect sizes on average, and the modest differences in estimates between Columns (2) and (4) are not statistically significant.

[28] We count an indicator as appearing in the index if it receives at least as much weight as it would in a KLK index; in this case that corresponds to a weight of $0.023 = \frac{1}{44}$.

[29] Let $r$ index 10% sample draws ($R = 100$), $h$ index indicator variables ($H = 44$), and $w_{hr}$ be the optimus weight for indicator $h$ in sample draw $r$. The average optimus weight for indicator variable $h$, applied to construct the "true" optimus index in the 100% sample, is: $\sum_{r=1}^{R} w_{hr} / \sum_{h=1}^{H} \sum_{r=1}^{R} w_{hr}$.

Table 6 also reports "effect sizes" and "index size" for an exhaustive PAP. The average PAP "effect size", reported in Column (2), corresponds to the average effect size for indicators that reject with the exhaustive PAP; when no indicator rejects (which occurs 42% of the time), the calculation includes the effect size for the most significant indicator. The average effect size for a PAP-rejected indicator is 0.21 standard deviations — more than double the optimus effect size and five times the overall average effect size. The average PAP "index size", or number of rejected indicators, is 1.3.[30] Finally, Column (4) reports the "true" average effect size (estimated on the 100% sample) for the indicators that reject in the exhaustive PAP. The average true effect size for these indicators is 0.13 standard deviations, or 38% less than the estimated average effect size. The discrepancy between the estimated effect size and the true effect size arises because the exhaustive PAP is underpowered and selects the most significant indicators for rejection, inflating the effect sizes (Ioannidis, 2008).[31] More generally, the PAP results in Table 6 highlight the difficulty in estimating and interpreting effect sizes with a PAP that tests many outcome variables: few indicators may be significant; those that are significant may feature large (true) effect sizes; and estimated effect sizes may be inflated without additional bias corrections (Andrews et al., 2021).

The analysis plan in Figure 4 specifies the optimus as a gate for the 44 indicator hypotheses — this is equivalent to executing an exhaustive PAP if and only if the optimus index rejects. The bottom row in Table 6 reports the average PAP power and "index size" when the optimus gates the PAP. In addition to rejecting the optimus index 71% of the time, the gated PAP rejects one or more indicators 51% of the time, which is only 7 percentage points lower than the ungated exhaustive PAP. Furthermore, the average number of rejections (conditional on rejecting anything) is virtually identical for the gated and ungated PAPs. In summary, adding an optimus gate to the exhaustive PAP comes at little cost — in 88% of cases in which the ungated exhaustive PAP would detect an effect on any individual outcome, the gated exhaustive PAP would also detect an effect on the same outcomes.

Appendix Table A8 reports the main results in Table 6 for the 8%, 12%, and 15% samples. Across all four sample sizes (8%, 10%, 12%, and 15%) the results are qualitatively similar: the optimus test has the highest power, followed by the KLK index, the exhaustive PAP, and the LMS omnibus test. Power for all tests increases with sample size, and the optimus's average effect

[30] To make the PAP result more comparable to the optimus and KLK index size figures, which are averaged across all 100 random samples, we left-censor the PAP "index size" at 1 in the 42% of random samples that reject nothing. If we set the PAP "index size" to 0 when nothing rejects, the average PAP "index size" is 0.9.

[31] Ironically, since the multiplicity-adjusted significance threshold is more stringent than the conventional significance threshold, the inflation bias can be even more extreme with an exhaustive PAP than the typical case of publication bias.

size and index size increase modestly with sample size, suggesting that larger samples allow the optimus to more precisely select variables for inclusion.

We also leverage the full dataset (i.e. 100% sample) to examine the frequency at which the optimus index includes indicators for which there is (approximately) a null effect. To determine this frequency, we first enumerate the outcomes that reject in the full dataset when controlling the false discovery rate at $q < 0.1$ (Benjamini et al., 2006); we find that there are 19 outcomes for which there is compelling evidence of a treatment effect. We then compute the average optimus weight assigned to each outcome across the 100 random 10% samples. We find that on average 81% of the weight in the optimus index gets assigned to variables for which there is strong evidence of a treatment effect in the full dataset. Only seven outcomes for which there is weak evidence of a treatment effect receive more than 1% average weight, with the most frequent "null" outcome (currently taking prescription medications) receiving 2.3% weight on average.[32] Thus, in expectation a supermajority of the optimus weight goes to outcomes with treatment effects, implying that in this case the optimus selects a broad index of variables that are generally affected by treatment. By comparison, the KLK index places 57% (25/44) of its weight on outcomes for which there is not strong evidence of a treatment effect.

Finally, we compare the performance of the optimus index and the KLK index when testing the three OHIE subfamily hypotheses: utilization-related outcomes, health-related outcomes, and financial outcomes. These estimates allow us to examine the performance of the different index tests in smaller families of hypotheses. Instead of testing a single all-outcome gate, we now test three subfamily indices in parallel. Appendix Table A9 reports average power, effect size, and index size for these three subfamilies using the 10% sample.[33] The tests are underpowered for all three subfamilies, in part because we now multiplicity adjust the *p*-values to reflect that we test the three subfamilies in parallel. Nevertheless, in each case the optimus outperforms the KLK index in terms of power. For example, for utilization-related outcomes the optimus achieves 36% power versus 21% power for the KLK index, while for financial-related outcomes the optimus achieves 23% power versus 17% power for the KLK index.[34] The results are similar if we instead consider power to pass an all-outcome gate and reject a subfamily index (Appendix Table A10), implying that subfamily indices rarely reject when the all-outcome index fails to reject. Furthermore, even

[32]This outcome has an unadjusted *p*-value of 0.10 and a FDR control *q*-value of 0.14 in the full dataset, suggesting that its null is more likely false than true.

[33]Appendix Table A7 reports the average weights received by each indicator when estimating a separate optimus index for each subfamily.

[34]The KLK index achieves only 2% power for the health-related outcomes. This is effectively the size of the test, since we multiplicity adjust the *p*-values for three parallel tests using the Romano-Wolf algorithm.

when defining three small, homogeneous families that each correspond to only one or two tables in the relevant OHIE PAP, we still find that the optimus outperforms a comparable KLK index.[35]

# 6 Conclusions and Recommendations

Analysis plans allow researchers to limit the rate of false discoveries through statistical adjustments for multiple inference. They do so at a cost: formal tests for multiple inference are only available for hypotheses which can be registered in an analysis plan, and these adjustments reduce power. If researchers fail to anticipate which indicators are most impacted by treatment, or can be measured with the least error, power concerns compound. These issues may lead to analysis plans which foreclose true discoveries in attempts to avoid false ones. In large part for this reason, Banerjee et al. (2020) emphasize the need for a central role of secondary evidence in academic research, even at the risk of promoting results based on type I errors.

In this paper, we suggest an alternate approach for analysis plans. Rather than restricting attention to the types of tests which can be easily anticipated, researchers can allow the data to inform the most powerful tests to be run. In a sense, specifying straightforward and easily anticipatable tests is an informal means of controlling false discoveries that becomes redundant when an analysis plan allows formal control. This fact allows researchers to specify data-driven analysis plans that can maximize the power of statistical tests. We propose the optimus gate as a method for doing so: by maximizing power among weighted index hypotheses and directing type I error to that high-power index test, researchers can be guaranteed a high power, easily interpretable test. We demonstrate in simulations and in two applications that this approach has substantial power advantages over other available approaches. The tradeoff for this power is a loss of control over which indices are being tested; researchers with strong priors over heterogeneity in treatment effects and who anticipate that a test based on a particular unweighted or weighted average treatment effect is of more inherent interest than tests based on alternate weighted average treatment effects will need to consider whether the gains in statistical power justify the loss of control.

The other challenge to the use of analysis plans to control false discoveries is complexity in specifying algorithmic approaches. The optimus gate approach not only maximizes power among available index hypotheses, it can also be straightforward to plan for, as a researcher need only

[35] We define three subfamilies that correspond to three key results tables in the Finkelstein et al. (2010) PAP: health care utilization (Tables P1 and U1), financial strain (Table P2), and health (Table P3). These subfamilies each contain eight, four, and seven indicators respectively. Appendix Table A11 reports average power, effect size, and index size for these three subfamilies using the 10% sample. For all three subfamilies, the optimus index has higher power than the KLK index.

categorize indicators into families. In many cases, particularly for researchers following the recommendations of Banerjee et al. (2020) in forming a preanalysis plan, these families may be relatively simple. For example, a straightforward gating structure would be to categorize indicators into first stage indicators which document whether a program was successfully implemented and second stage indicators which indicate whether a successfully implemented program influenced important economic outcomes. A simple preanalysis plan may set optimus gates for the first stage hypothesis followed by the second, and in doing so would guarantee that researchers find a powerful test of whether the program was implemented and whether it impacted outcomes in a way that maintains correct size on each test. Researchers with more time and stronger priors over which potential categories of effects a program may have may add a third set of gates which test several of these categories in parallel; an additional advantage of the optimus gate approach is that such a plan would not harm statistical power on the first two primary tests.

While the optimus approach will not be preferred in all scenarios, it can be attractive in most scenarios in which researchers anticipate testing large numbers of outcomes with heterogeneous effect sizes in a limited sample. In our experience this scenario is well-represented among field experiments.

# 7 References

# References


Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. Journal of the American Statistical Association, 103(484):1481–1495.

Anderson, M. L. and Magruder, J. (2017). Split-sample strategies for avoiding false discoveries. NBER Working Paper No 23544.

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. American Economic Review, 109(8):2766–94.

Andrews, I., McCloskey, A., and Kitagawa, T. (2021). Inference on winners. Working Paper, Harvard University.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., and Sautmann, A. (2020).

In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. NBER Working Paper No 26993.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3):491–507.

Bretz, F., Maurer, W., and Hommel, G. (2011). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. Statistics in medicine, 30(13):1489–1501.

Casey, K., Glennerster, R., and Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. The Quarterly Journal of Economics, 127(4):1755–1812.

Chernozukhov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018). Generic machine learning on heterogeneous treatment effects in randomized experiments. NBER Working Paper No 24678.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. AEA Papers and Proceedings, 110:42–48.

Dallow, N. S., Leonov, S. L., and Roger, J. H. (2008). Practical usage of o'brien's ols and gls statistics in clinical trials. Pharmaceutical Statistics, 7(1):53–68.

De Angelis, C., Drazen, J., Frizelle, F., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P., Schroeder, T., Sox, H., and Van Der Weyden, M. (2004). Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors.

Dmitrienko, A. and Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry, 6(3):171–180.

Dmitrienko, A., Wiens, B. L., Tamhane, A. C., and Wang, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. Statistics in medicine, 26(12):2465–2478.

Fafchamps, M. and Labonne, J. (2017). Using Split Samples to Improve Inference about Causal Effects. Working Paper, Stanford University.

Finkelstein, A., Taubman, S., Allen, H., Gruber, J., Newhouse, J. P., Wright, B., Baicker, K., and Group, T. O. H. S. (2010). The short-run impact of extending public health insurance to low

income adults: evidence from the first year of the oregon medicaid experiment. Analysis Plan, NBER.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, T. O. H. S. (2012). The oregon health insurance experiment: Evidence from the first year. Quarterly Journal of Economics, 127:1057–1106.

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

Gerber, A. and Malhotra, N. (2008). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. Quarterly Journal of Political Science, 3(3):313–326.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer New York.

Horton, R. and Smith, R. (1999). Time to register randomised trials. BMJ, 319(7214):865–866.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, pages 640–648.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. Econometrica, 75(1):83–119.

Ludwig, J., Mullainathan, S., and Spiess, J. (2019). Machine-learning tests for effects on multiple outcomes. Mimeo, Harvard.

Neumark, D. (2001). The employment effects of minimum wages: Evidence from a prespecified research design the employment effects of minimum wages. Industrial Relations: A Journal of Economy and Society, 40(1):121–144.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. Biometrics, 40(4):1079–1087.

Olken, B. A. (2015). Promises and Perils of Pre-analysis Plans. Journal of Economic Perspectives, 29(3):61–80.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. Biometrics, 43(3):487–498.

Romano, J. P. and DiCiccio, C. (2019). Multiple data splitting for testing. Department of Statistics, Stanford University.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2010). Hypothesis testing in econometrics. Annual Review of Economics, 2(1):75–104.

Romano, J. P. and Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. Econometrica, 73(4):1237–1282.

Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials. Journal of Clinical Oncology, 4(10):1529–1541.

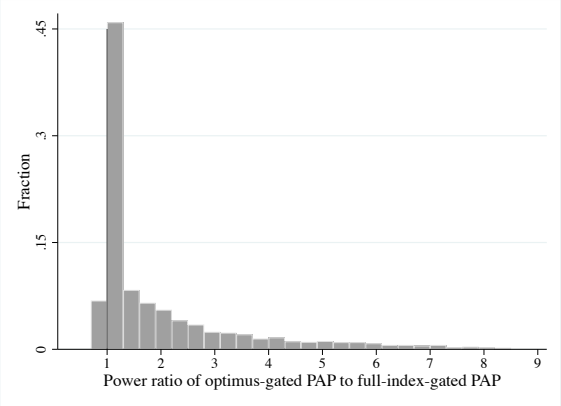
Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa. Journal of the American Statistical Association, 54(285):30–34.

Yong, E. (2012). Replication studies: Bad copy. Nature, 485(7398):298–300.

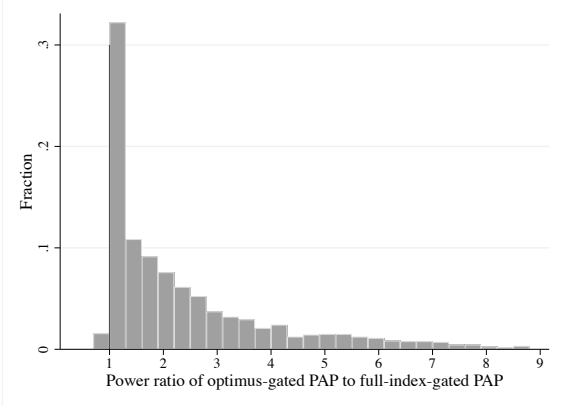
# Figures and Tables

Figure 1: Distribution of Relative Power of Optimus versus KLK Index Gatekeeping Strategy

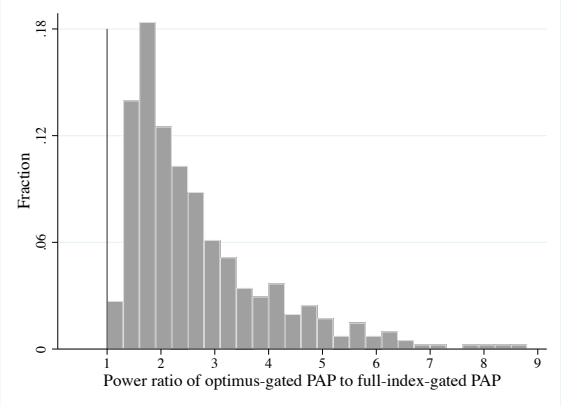
(a) All parameter combinations



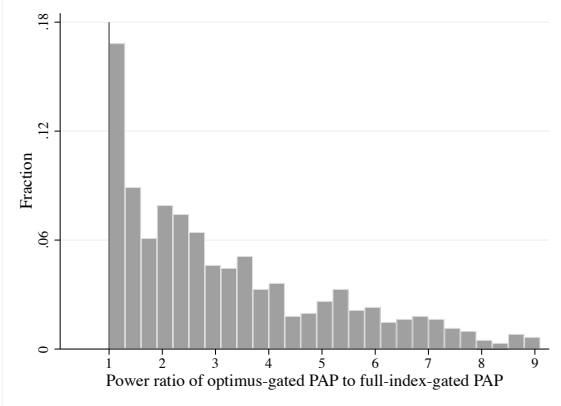

(b) More relevant parameter combinations



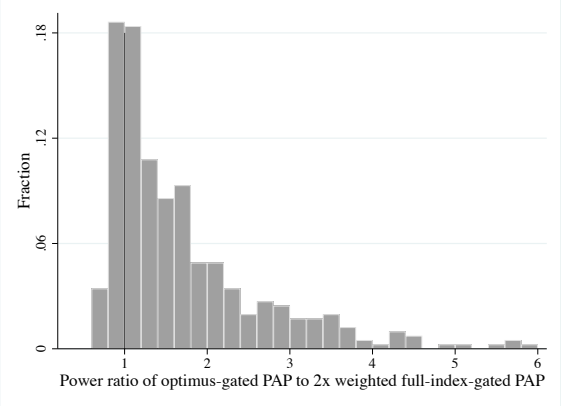

(c) Small families (equal weight)



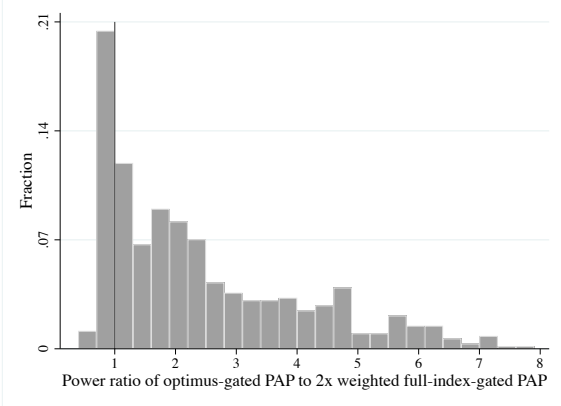

(d) Large families (equal weight)




(e) Small families (KLK 2x weight)



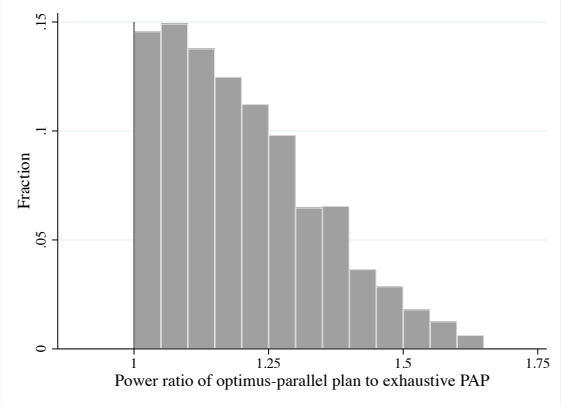

(f) Large families (KLK 2x weight)



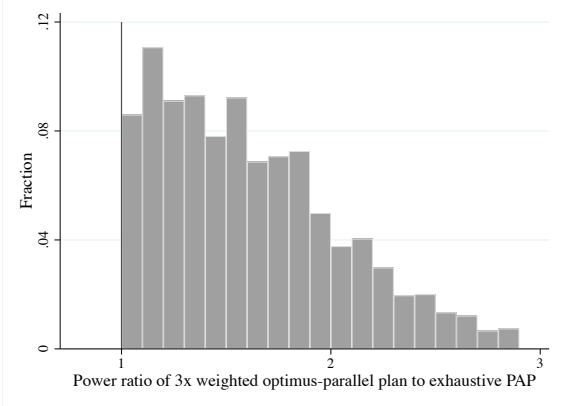

*Notes:* Panels (a) and (b) correspond to Columns (1) and (2) of Table 2 respectively, Panels (c) and (e) correspond to Column (3), and Panels (d) and (f) correspond to Column (4).

Figure 2: Distribution of Relative Power of Optimus Plans versus Exhaustive PAP

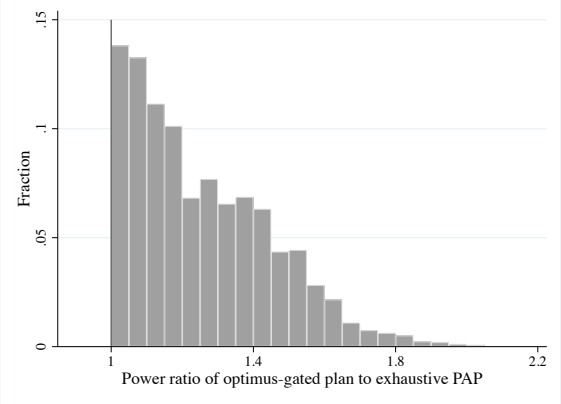
(a) Parallel optimus plan (index weight = 1)



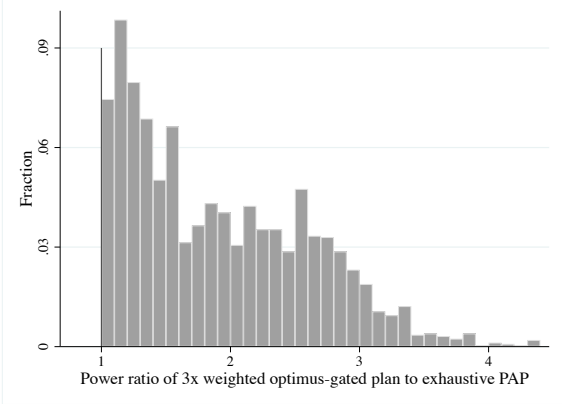

(b) Parallel optimus plan (index weight = 3)




(c) Gated optimus plan (index weight = 1)




(d) Gated optimus plan (index weight = 3)




*Notes:* All panels correspond to Column (1) of Table 2.

Figure 3: Optimus Gate analysis plan for Casey et al. (2012)

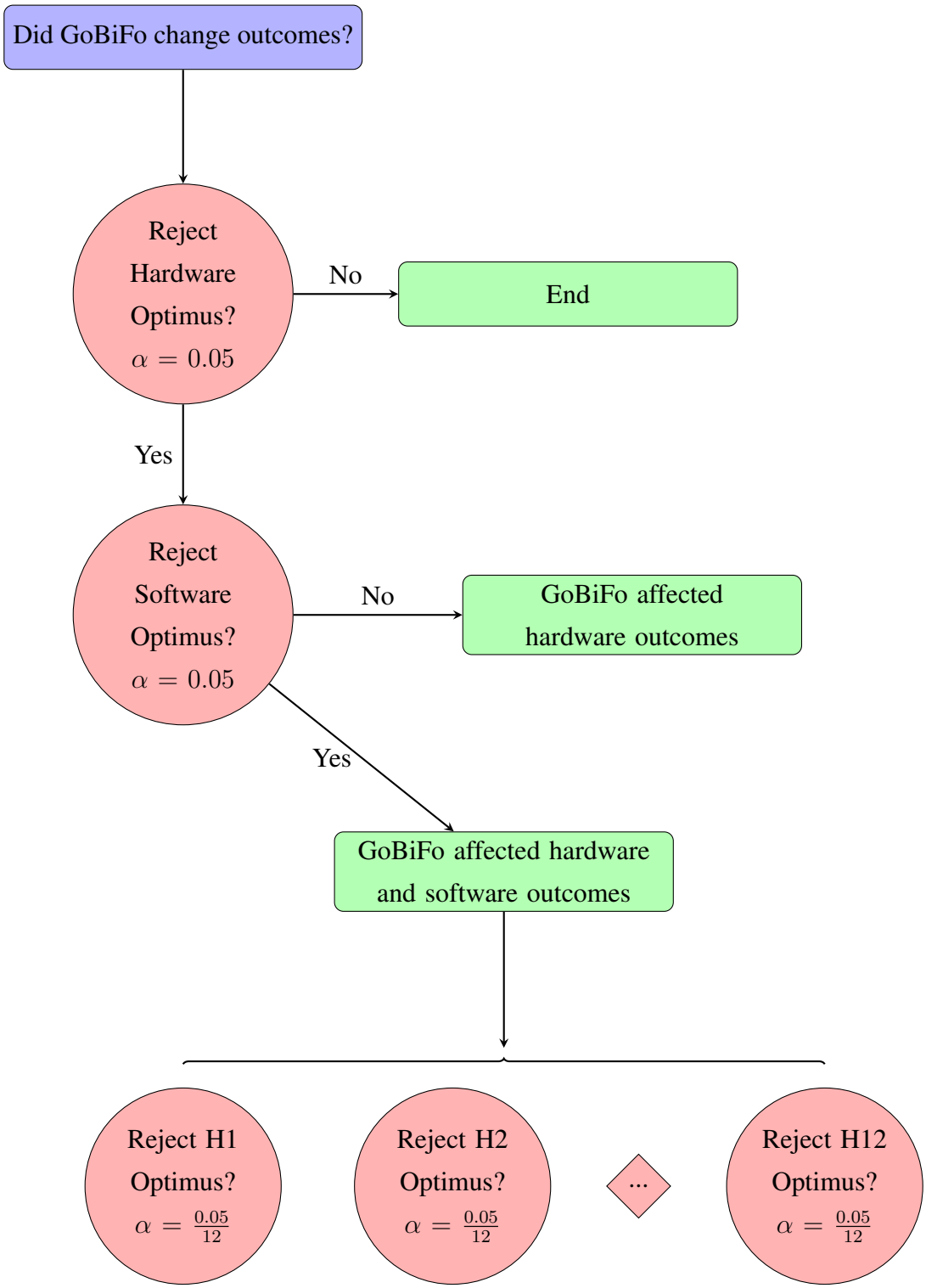


Figure 4: Optimus Gate analysis plan for Finkelstein et al. (2012)

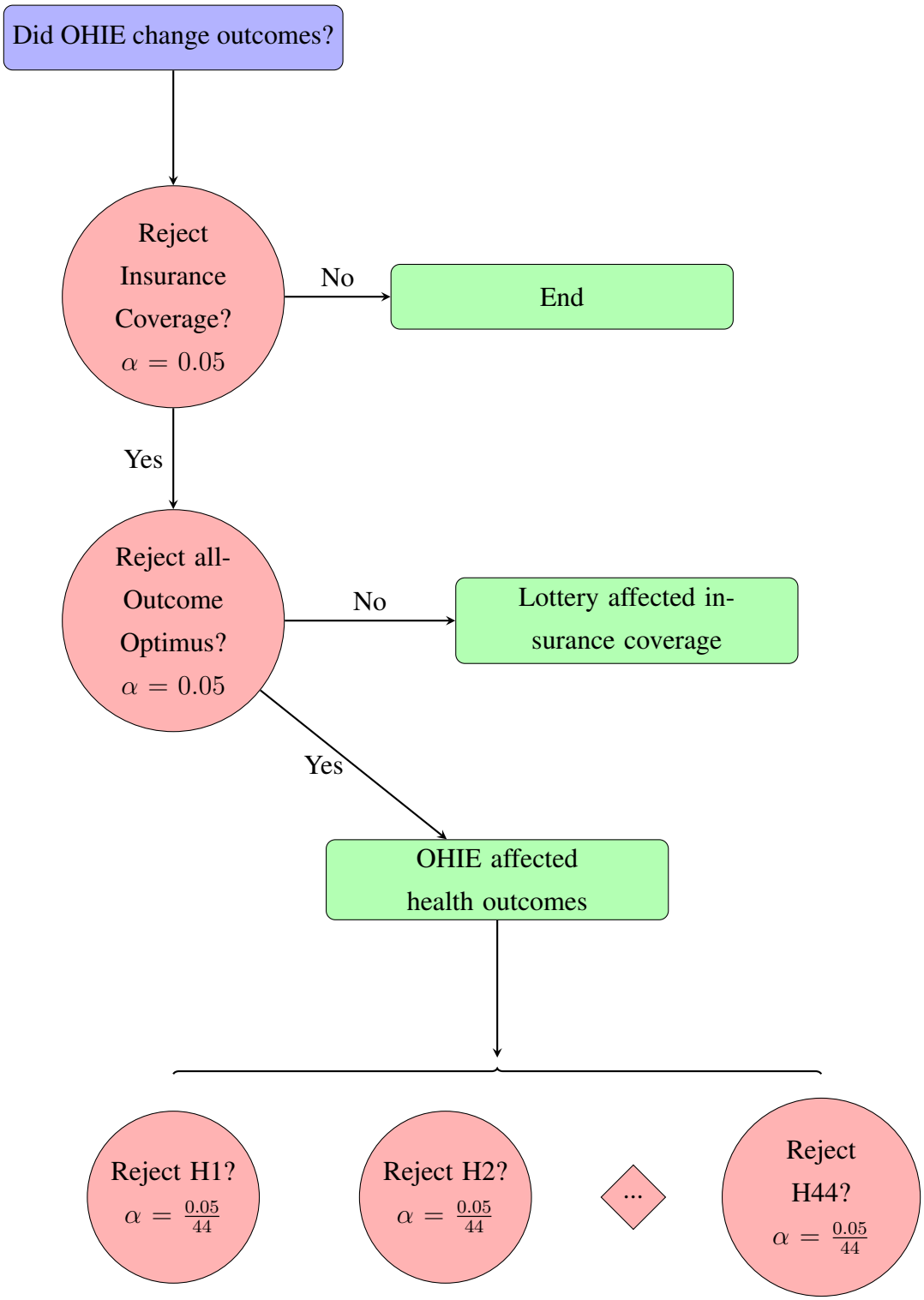


Table 1: Simulation Parameter Values

| Parameter | Values | Mean | Std Dev |
|---|---|---|---|
| Average "effect size" ($E[t_h \mid \beta_h \neq 0]$) | 1.5, 2.0, 2.5, 3.0, 4.0 | 2.6 | 0.9 |
| Total hypotheses ($H$) | 10, 20, 50, 100, 200 | 76 | 69 |
| Share false ($H_1/H$) | 0.1, 0.2, 0.5, 1.0 | 0.45 | 0.35 |
| Correlation between outcomes ($\rho$) | 0, 0.1, 0.2, 0.5, 0.7 | 0.35 | 0.25 |
| Share of outcomes correlated ($r$) | 0.2, 0.5, 1.0 | 0.6 | 0.34 |

Table 2: Relative Power of Optimus Index v. KLK Index

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Index test only | 1.77 | 2.24 | 2.55 | 3.28 | 3.48 |
| Index test in parallel w/PAP | 3.64 | 5.98 | 6.07 | 15.9 | 13.8 |
| Optimus index size | 17.1 | 14.5 | 4.9 | 18.5 | 13.8 |
| *Parameter restrictions:* | | | | | |
| Total hypotheses ($H$) | | | $\leq 20$ | $\geq 50$ | |
| Share false ($H_1/H$) | | $\leq 0.5$ | $\leq 0.2$ | $\leq 0.2$ | $0.1$ |
| Average effect size ($\mu_t$) | | $\leq 3.0$ | $\leq 3.0$ | $\leq 3.0$ | $\leq 2.5$ |
| Combinations | 2,600 | 1,560 | 416 | 624 | 390 |

Notes: Each cell reports the geometric mean power ratio of an optimus index to a KLK index. The index is tested by itself (first row) or in parallel with an exhaustive PAP (second row).

Table 3: Relative Power of Optimus-Gated Plans v. Optimus-Parallel Plans

| *Optimus-index weight:* | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| 1.0 | 1.04 | 1.06 | 1.05 | 1.08 | 1.07 |
| 2.0 | 1.11 | 1.17 | 1.24 | 1.19 | 1.30 |
| 3.0 | 1.15 | 1.24 | 1.37 | 1.27 | 1.46 |
| 4.0 | 1.18 | 1.29 | 1.47 | 1.32 | 1.59 |
| 5.0 | 1.21 | 1.33 | 1.54 | 1.37 | 1.69 |
| *Parameter restrictions:* | | | | | |
| Total hypotheses ($H$) | | | $\leq 20$ | $\geq 50$ | |
| Share false ($H_1/H$) | | $\leq 0.5$ | $\leq 0.2$ | $\leq 0.2$ | $0.1$ |
| Average effect size ($\mu_t$) | | $\leq 3.0$ | $\leq 3.0$ | $\leq 3.0$ | $\leq 2.5$ |
| Combinations | 2,600 | 1,560 | 416 | 624 | 390 |

Notes: Each cell reports, for a given optimus-index weight, the geometric mean power ratio of an optimus-index gated exhaustive PAP to an exhaustive PAP that tests the optimus index in parallel with the other hypotheses.

Table 4: GoBiFo Results: Individual Hypotheses

| | (1) KLK Index | (2) Optimus | (3) KLK Index Size | (4) Optimus Index Size |
|---|---|---|---|---|
| *Hardware* | | | | |
| H1: GoBiFo Program Implementation | 0.695 [0.00] | 1.633 [0.000] | 7 | 2.5 |
| H2: Participation in GoBiFo improves the quality of local public service infrastructure | 0.206 [0.00] | 0.494 [0.000] | 18 | 6.8 |
| H3: Participation in GoBiFo improves General Economic Welfare | 0.362 [0.00] | 1.864 [0.000] | 15 | 2.9 |
| *Software* | | | | |
| H4: Participation in GoBiFo increases collective action and contributions to public goods. | -0.001 [1] | 0.288 [.15] | 11 | 2.3 |
| H5: GoBiFo increases inclusion and participation in community planning and implementation | -0.002 [1] | -0.004 [.962] | 46 | 0.9 |
| H6: GoBiFo challenges local systems of authority | 0.052 [.74] | 0.183 [.063] | 25 | 6.4 |
| H7: Participation in GoBiFo increases trust | 0.036 [1] | 0.043 [.938] | 12 | 1.7 |
| H8: Participation in GoBiFo builds and strengthens community groups and networks | 0.027 [1] | 0.209 [.35] | 15 | 2.6 |
| H9: Participation in GoBiFo increases access to information about local governance | 0.01 [1] | 0.043 [.925] | 15 | 2.6 |
| H10: GoBiFo increases public participation in local governance | -0.028 [1] | -0.041 [.95] | 14 | 0.8 |
| H11: By increasing trust, GoBiFo reduces crime and conflict in the community. | 0.014 [1] | 0.139 [.681] | 8 | 2.2 |
| H12: GoBiFo changes political and social attitudes | 0.035 [1] | 0.062 [.887] | 9 | 2.5 |

Notes: Brackets contain Romano-Wolf $p$-values that control FWER across all 12 hypotheses, computed based on 80 permutations of treatment under the null hypothesis. Optimus index $p$-values represent the median RW $p$-value across 200 sets of 5-fold assignments (computed based on 80 permutations of treatment under the null hypothesis per set of fold assignments).

Table 5: GoBiFo Results: Gating Hypotheses

| | (1) KLK Index | (2) Optimus | (3) LMS Omnibus | (4) KLK Index Size | (5) Optimus Index Size |
|---|---|---|---|---|---|
| Hardware | 0.292 | 1.191 | | 39 | 11.3 |
| | [0.000] | [0.000] | [0.000] | | |
| Software | 0.014 | 0.151 | | 144 | 22.8 |
| | [.473] | [0.000] | [.338] | | |
| Software (no Community Farm) | 0.011 | 0.123 | | 143 | 22.1 |
| | [.578] | [.013] | [.419] | | |

Notes: Brackets contain $p$-values. Optimus and LMS omnibus $p$-values represent the median $p$-value across 200 sets of 5-fold assignments, computed based on 80 permutations of treatment under the null hypothesis per set of fold assignments.

Table 6: OHIE Results: Gating and Individual Indicator Hypotheses

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *Test:* | Power | Effect Size | Index Size | True Effect Size |
| Optimus | 71% | 0.092 | 9.0 | 0.099 |
| KLK Index | 58% | 0.038 | 44 | 0.039 |
| LMS Omnibus | 38% | | | |
| Exhaustive PAP | 58% | $0.213^{+}$ | $1.3^{++}$ | 0.131 |
| PAP (post optimus gate) | 51% | $0.214^{+}$ | $1.3^{++}$ | 0.131 |

Notes: Results in Columns (1) – (3) represent averages across 100 random 10% samples of the OHIE data. Power denotes power to reject the sharp null hypothesis for at least one indicator. True effect size represents the estimated effect in the full (100% sample) OHIE dataset, with indicators weighted using average weights underlying Column (2).

+ Average effect size for indicators rejected by the PAP (when no indicator rejects, calculation includes the most significant indicator).

++ Average number of indicators rejected, left-censored at 1.

# Appendix
## Not For Print Publication

## A1 Empirical Distribution of *t*-statistics

Our sample consists of papers on field experiments published from 2013 to 2015 in a set of ten general-interest economics journals: *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Review*, *Econometrica*, *Economic Journal*, *Journal of the European Economic Association*, *Journal of Political Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics and Statistics*. We defined a paper as involving field experiments if it mentioned "field experiment" in its abstract or listed JEL Code C93. These criteria generated a sample of 61 papers. Using this sample we recorded the *t*-statistic for each paper's featured result. The median *t*-statistic was 2.6, the 10th percentile *t*-statistic was 1.7, and the 90th percentile *t*-statistic was 7.0. Due to the likelihood of publication bias and *p*-hacking (Franco et al. 2014), we interpret this distribution as an overestimate of the *ex ante t*-statistic distribution that a researcher should expect when beginning a typical field experiment. Nevertheless, the results imply that most researchers should (at best) expect statistical power that corresponds to mean "effect sizes" (i.e. *t*-statistics) of 2.0 to 3.0 in our power simulations, and we focus our discussion on effect sizes in this range.

## A2 Mathematical Proofs

Let $\bar{y}_{i\mathbf{w}_g} = \mathbf{w}_g' \mathbf{y}_{ig}$ and $\hat{\beta}_{\mathbf{w}_g}$ be the coefficient associated with a regression of $\bar{y}_{i\mathbf{w}_g}$ on treatment (represented by Equation (6)). Let all outcomes be standardized to have unit variance and let the elements of $\mathbf{w}_g$ sum to one.

### A2.1 Index Power

**Lemma 1.** *Let* $\beta_{\mathbf{g}}{}' = (\beta_{1g}\ \beta_{2g}\ ...\ \beta_{|\mathcal{H}_g|g})$. *For a given* $\mathbf{w}_g$, *let* $\beta_{\mathbf{w}_g} = \beta_{\mathbf{g}}{}'\mathbf{w}_g$. *A regression of* $\bar{y}_{i\mathbf{w}_g}$ *on treatment estimates* $\beta_{\mathbf{w}_g}$.

Proof: Let $\tilde{\mathbf{T}}$ be a demeaned $N \times 1$ vector of treatment assignments, $\mathbf{y}$ any $N \times 1$ vector of outcomes, and $\mathbf{Y_g}$ the $N \times |\mathcal{H}_g|$ matrix of outcomes in group $g$. A regression of $y$ on treatment

recovers $\hat{\beta} = (\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}(\tilde{\mathbf{T}}'\mathbf{y})$. Thus a regression of $\bar{y}_{i\mathbf{w}_g}$ on treatment yields

$$\begin{aligned}(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}(\tilde{\mathbf{T}}'\bar{\mathbf{y}}_{\mathbf{w_g}}) &= (\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}(\tilde{\mathbf{T}}'\mathbf{Y}_g\mathbf{w}_g) \\ &= (\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}(\tilde{\mathbf{T}}'\mathbf{Y}_g)\mathbf{w}_g = \hat{\beta}'_{\mathbf{g}}\mathbf{w}_g\end{aligned}$$

For completeness note that $\hat{\beta}_{hg}$ is a consistent estimator of $\beta_{hg}$, so the regression coefficient is a consistent estimator of $\beta_{\mathbf{w}_g}$.

**Lemma 2.** *For a given $\mathbf{w}_g$, let $\hat{\beta}_{\mathbf{w}_g}$ be the coefficient associated with a regression of $\bar{y}_{i\mathbf{w}_g}$ on treatment. The variance of $\hat{\beta}_{\mathbf{w}_g}$ is $\mathbf{w}'_g\Sigma_g\mathbf{w}_g$, where $\Sigma_g$ represents the covariance matrix for $\hat{\beta}_{\mathbf{g}}$.*

Proof: From Lemma 1, $\hat{\beta}_{\mathbf{w}_g} = \hat{\beta}'_{\mathbf{g}}\mathbf{w}_g$. Then

$$\mathrm{V}(\hat{\beta}_{\mathbf{w}_g}) = \mathrm{V}(\hat{\beta}'_{\mathbf{g}}\mathbf{w}_g) = \mathbf{w}'_g\mathrm{V}(\hat{\beta}_{\mathbf{g}})\mathbf{w}_g = \mathbf{w}'_g\Sigma_g\mathbf{w}_g$$

**Proposition 1.** *Consider an index $\bar{y}_{i\mathbf{w}_g} = \mathbf{w}'_g\mathbf{y}_{ig}$. Let $\hat{\beta}_{\mathbf{w}_g}$ be the regression coefficient from estimating Equation (6) and let $\sigma_{\hat{\beta}_{\mathbf{w}_g}} = \sqrt{V(\hat{\beta}_{\mathbf{w}_g})}$. A one-sided test of $\beta_{\mathbf{w}_g} = 0$ based on $\hat{\beta}_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}}$ with critical value $\Phi^{-1}(1-\alpha)$ has power $\Phi(\frac{\beta_{\mathbf{g}'}\mathbf{w}_g}{\sqrt{\mathbf{w}'_g\Sigma_g\mathbf{w}_g}} + \Phi^{-1}(\alpha))$.*

Proof: Following convention, let $\hat{\beta}_{\mathbf{w}_g}$ be distributed $N(\beta_{\mathbf{w}_g}, \mathrm{V}(\hat{\beta}_{\mathbf{w}_g}))$. Applying Lemmas 1 and 2,

$$\begin{aligned}&\mathbb{P}(\hat{\beta}_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}} > \Phi^{-1}(1-\alpha)) \\ &= \mathbb{P}((\hat{\beta}_{\mathbf{w}_g} - \beta_{\mathbf{w}_g})/\sigma_{\hat{\beta}_{\mathbf{w}_g}} > -\beta_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}} - \Phi^{-1}(\alpha)) \\ &= \mathbb{P}((\beta_{\mathbf{w}_g} - \hat{\beta}_{\mathbf{w}_g})/\sigma_{\hat{\beta}_{\mathbf{w}_g}} < \beta_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}} + \Phi^{-1}(\alpha)) \\ &= \Phi(\beta_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}} + \Phi^{-1}(\alpha))) \\ &= \Phi(\tfrac{\beta_{\mathbf{g}'}\mathbf{w}_g}{\sqrt{\mathbf{w}'_g\Sigma_g\mathbf{w}_g}} + \Phi^{-1}(\alpha)).\end{aligned}$$

**Corollary 1.1.** *Consider a flatly-weighted KLK index for family $g$, $\bar{y}_i = \frac{1}{|\mathcal{H}_g|}\sum_{h\in\mathcal{H}_g} y_{ihg}$. Let $\hat{\beta}_{\bar{y}}$ be the regression coefficient from a regression of $\bar{y}_i$ on treatment. Suppose $\beta_{hg} = \beta\ \forall\ h \in \mathcal{H}_g$. Let $\sigma_{\hat{\beta}_{\bar{y}}} = \sqrt{V(\hat{\beta}_{\bar{y}})}$ and $\sigma_{\hat{\beta}_{hg}} = \sqrt{V(\hat{\beta}_{hg})}$. A test of $\beta > 0$ based on $\hat{\beta}_{\bar{y}}/\sigma_{\hat{\beta}_{\bar{y}}}$ with critical value $\Phi^{-1}(1-\alpha)$ is weakly more powerful than a test of $\beta > 0$ based on $\hat{\beta}_{hg}/\sigma_{\hat{\beta}_{hg}}$ with the same critical value.*

Proof: Following convention, let $\hat{\beta}_{hg}$ be distributed $N(\beta, \sigma^2_{\hat{\beta}_{hg}})$. Note that $\bar{y}_i = \frac{\mathbf{1}'\mathbf{y}_{ig}}{|\mathcal{H}_g|}$, so $\mathbf{w}_g$ for the KLK index is $\frac{\mathbf{1}}{|\mathcal{H}_g|}$. Applying Proposition 1, the test based on $\hat{\beta}_{\bar{y}}/\sigma_{\hat{\beta}_{\bar{y}}}$ has power $\Phi(\frac{\beta_{\mathbf{g}'}\mathbf{1}}{\sqrt{\mathbf{1}'\Sigma_g\mathbf{1}}} +$

$\Phi^{-1}(\alpha))$. In the special case in which $\mathbf{w}_g$ contains a single non-zero element the test has power $\Phi(\frac{\beta_{hg}}{\sigma_{\hat{\beta}_{hg}}} + \Phi^{-1}(\alpha)) \; \forall \; h \in \mathcal{H}_g$. Then

$$\begin{aligned}
&\mathbb{P}(\hat{\beta}_{\bar{y}}/\sigma_{\hat{\beta}_{\bar{y}}} > \Phi^{-1}(1-\alpha)) \\
&\quad = \Phi(\frac{\beta_{\mathbf{g}'\mathbf{1}}}{\sqrt{\mathbf{1}'\Sigma_g\mathbf{1}}} + \Phi^{-1}(\alpha)) \\
&\quad = \Phi(\frac{|\mathcal{H}_g|\beta}{\sqrt{\mathbf{1}'\Sigma_g\mathbf{1}}} + \Phi^{-1}(\alpha)) \\
&\quad \geq \Phi(\frac{|\mathcal{H}_g|\beta}{\sqrt{|\mathcal{H}_g|^2\sigma^2_{\hat{\beta}_{hg}}}} + \Phi^{-1}(\alpha)) \\
&\quad = \Phi(\frac{\beta_{hg}}{\sigma_{\hat{\beta}_{hg}}} + \Phi^{-1}(\alpha)) \\
&= \mathbb{P}(\hat{\beta}_{hg}/\sigma_{\hat{\beta}_{hg}} > \Phi^{-1}(1-\alpha)).
\end{aligned}$$

**Corollary 1.2.** *Consider a KLK index for family $g$, $\bar{y}_i = \frac{1}{|\mathcal{H}_g|}\sum_{h\in\mathcal{H}_g} y_{ihg}$, and an alternative index $\mathbf{w}_g'\mathbf{y}_{ig}$ with $\mathbf{w}_g \not\propto \mathbf{1}$. Suppose $\beta_{hg} = \beta \; \forall \; h \in \mathcal{H}_g$ and $\Sigma_g = \sigma^2_{\hat{\beta}_{hg}}[(1-\rho)\cdot I + \rho\cdot\mathbf{1}\mathbf{1}']$ for $-1 \leq \rho \leq 1$. Let $\sigma_{\hat{\beta}_{\bar{y}}} = \sqrt{V(\hat{\beta}_{\bar{y}})}$ and $\sigma_{\hat{\beta}_{\mathbf{w}_g}} = \sqrt{V(\hat{\beta}_{\mathbf{w}_g})}$. A test of $\beta > 0$ based on $\hat{\beta}_{\bar{y}}/\sigma_{\hat{\beta}_{\bar{y}}}$ with critical value $\Phi^{-1}(1-\alpha)$ is weakly more powerful than a test of $\beta > 0$ based on $\hat{\beta}_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}}$ with the same critical value.*

Proof: Applying Proposition 1, a test based on $\hat{\beta}_{\mathbf{w}_g}/\sigma_{\hat{\beta}_{\mathbf{w}_g}}$ has power $\Phi(\frac{\beta_{\mathbf{g}'\mathbf{w}_g}}{\sqrt{\mathbf{w}_g'\Sigma_g\mathbf{w}_g}} + \Phi^{-1}(\alpha))$. But $\beta_{\mathbf{g}}{}'\mathbf{w}_g = \beta$ since $\mathbf{w}_g'\mathbf{1} = 1$, so maximizing power is equivalent to minimizing $a(\mathbf{w}_g) = \mathbf{w}_g'\Sigma_g\mathbf{w}_g$. The gradient of $a(\mathbf{w}_g)$ is $2\Sigma_g\mathbf{w}_g$ and the Hessian is $2\Sigma_g$; thus $\partial a/\partial w_{hg} = 2(w_{hg} + \rho\sum_{j\neq h} w_{jg}) \; \forall \; h \in \mathcal{H}_g$. The optimal weights are therefore identical across hypotheses, with $\mathbf{w}_g^* = \mathbf{1}/|\mathcal{H}_g|$ given the constraint that weights sum to one. These optimal weights yield the KLK index $\bar{y}_i$.

## A2.2 Full Sample and *K*-fold Optimus Results

Consider a "full sample" optimus test that does not use cross-validation. We start with a group-level stacked version of the statistical model in Equation (1)

$$\mathbf{y}_{ig} = \beta_{\mathbf{g}} T_i + \varepsilon_{ig} \tag{10}$$

with $\mathbf{y}_{ig}$, $\beta_{\mathbf{g}}$, and $T_i$ as defined previously, and $\varepsilon_{ig}$ as an i.i.d. (across $i$ ) $|\mathcal{H}_g| \times 1$ column vector containing errors for all hypotheses $h \in \mathcal{H}_g$. After estimating Equation (10) we identify the vector

of weights $\omega_g$ that solves

$$\omega_g = \text{argmax}_{\mathbf{w}_g} \Phi\left(\frac{\hat{\beta}'_{\mathbf{g}} \mathbf{w}_g}{\sqrt{\mathbf{w}_g \widehat{\Sigma}_g \mathbf{w}_g}} + \Phi^{-1}(\alpha)\right)$$

**Proposition 2.** *Let* $\omega_g = argmax_{\mathbf{w}_g} \Phi\left(\frac{\hat{\beta}'_{\mathbf{g}} \mathbf{w}_g}{\sqrt{\mathbf{w}_g \widehat{\Sigma}_g \mathbf{w}_g}} + \Phi^{-1}(\alpha)\right)$*, where* $\hat{\beta}_{\mathbf{g}}$ *and* $\widehat{\Sigma}_g$ *are sample estimates of* $\beta_{\mathbf{g}}$ *and* $\Sigma_g$*. Let* $\beta_{\omega_g} = E[\beta_{\mathbf{g}}{}' \omega_g \mid \omega_g]$*. Consider an index* $\bar{y}_{i\omega_g} = \omega'_g \mathbf{y}_{ig}$*. A regression of* $\bar{y}_{i\omega_g}$ *on* $T_i$ *yields a biased estimate of* $\beta_{\omega_g}$*.*

Proof: The use of the sample estimators $\hat{\beta}_{\mathbf{g}}$ and $\widehat{\Sigma}_g$ implies that $\omega_g$ is a function of $T_i$ and $\varepsilon_{ig}$ $\forall$ $i$. In particular, the solution maximizes a generally increasing function of $\hat{\beta}'_{\mathbf{g}} \mathbf{w}_g = (\beta_{\mathbf{g}}{}' + (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\varepsilon_g)\mathbf{w}_g = \beta_{\mathbf{w}_g} + (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\varepsilon_g \mathbf{w}_g$, where $\varepsilon_g$ is a $N \times H$ matrix containing $\varepsilon_{ig}$ $\forall$ $i$. To highlight this dependency we write $\omega_g(\mathbf{T}, \varepsilon_g)$. In the regression

$$\bar{y}_{i\omega_g} = \beta_{\omega_g} T_i + \omega_g(\mathbf{T}, \varepsilon_g)' \varepsilon_{ig}$$

it is therefore generally the case that

$$E[\omega_g(\mathbf{T}, \varepsilon_g)' \varepsilon_{ig} \mid T_i] \neq 0$$

Developing the optimus test using the full sample thus generates a biased estimator of $\beta_{\omega_g}$, even when defining the target parameter to be conditional on the estimated optimus weights.

**Proposition 3.** *Randomly assign* $N$ *observations to* $K$ *folds. For each fold* $k$*, compute weights* $\omega_{-k,g} = argmax_{\mathbf{w}_{-k,g}} \Phi\left(\frac{\hat{\beta}'_{-\mathbf{k},\mathbf{g}} \mathbf{w}_{-k,g}}{\sqrt{\mathbf{w}_{-k,g} \widehat{\Sigma}_{-k,g} \mathbf{w}_{-k,g}}} + \Phi^{-1}(\alpha)\right)$*, where* $\hat{\beta}_{-\mathbf{k},\mathbf{g}}$ *and* $\widehat{\Sigma}_{-k,g}$ *are estimates of* $\beta_{\mathbf{g}}$ *and* $\Sigma_g$ *using all observations not in fold* $k$*. Let* $\tilde{\mathbf{T}}$ *be a demeaned* $N \times 1$ *vector of treatment assignments and* $\tilde{\mathbf{Y}}_g$ *be a* $N \times 1$ *vector of weighted outcomes, with element* $i$ *equal to* $\omega'_{-k,g} \mathbf{y}_{ig}$*. The* $K$*-fold optimus estimator* $(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\tilde{\mathbf{Y}}_g$ *is unbiased for* $E[\beta_{\mathbf{g}}{}' \omega_{-k,g}]$*.*

Proof: Let $p$ represent the proportion of treated observations ($T_i = 1$); thus $\sum_i (T_i - \bar{T})^2 = Np(1-p)$. Randomly assign $N$ observations to $K$ folds, stratifying folds on treatment $T_i$. Without loss of generality, assume the fold index $k$ weakly increases with $i$; we may write $k_i$ to refer to the fold containing observation $i$. For each fold $k$, the optimus is derived from the other $K - 1$ folds, with weights given by

$$\omega_{-k,g} = \text{argmax}_{\mathbf{w}_{-k,g}} \Phi\left(\frac{\hat{\beta}'_{-k,g} \mathbf{w}_{-k,g}}{\sqrt{\mathbf{w}_{-k,g} \widehat{\Sigma}_{-k,g} \mathbf{w}_{-k,g}}} + \Phi^{-1}(\alpha)\right)$$

where $\hat{\beta}_{-\mathbf{k},\mathbf{g}}$ and $\widehat{\Sigma}_{-k,g}$ are estimates of $\beta_{\mathbf{g}}$ and $\Sigma_g$ computed using all observations not in fold $k$. Let $\tilde{\mathbf{T}}$ be a demeaned $N \times 1$ vector of treatment assignments and let

$$\tilde{\mathbf{Y}}_g = \begin{bmatrix} \omega'_{-1,g}\mathbf{y}_{1g} \\ \omega'_{-1,g}\mathbf{y}_{2g} \\ ... \\ \omega'_{-K,g}\mathbf{y}_{Ng} \end{bmatrix}$$

The $K$-fold optimus estimator is $(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\tilde{\mathbf{Y}}_g$. To show unbiasedness first note that $E[\tilde{T}_i^2\omega'_{-k_i,g}\beta_{\mathbf{g}}] = E[\tilde{T}_i^2]E[\omega'_{-k_i,g}\beta_{\mathbf{g}}]$ and $E[\tilde{T}_i\omega'_{-k_i,g}\varepsilon_{ig}] = 0$ because $\omega_{-k_i,g}$ is a function of $\mathbf{T}_{-k_i}$ and $\varepsilon_{-k_i,g}$ (i.e. data from folds not containing $i$), and $\varepsilon_{ig}, \varepsilon_{jg}, T_i$, and $T_j$ are jointly independent $\forall\, i, j \ni i \neq j$. Then

$$\begin{aligned}
&E[(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\tilde{\mathbf{Y}}_g] \\
&= E[\frac{\sum_i \tilde{T}_i\omega'_{-k_i,g}\mathbf{y}_{ig}}{\sum_i \tilde{T}_i^2}] \\
&= E[\frac{\sum_i \tilde{T}_i\omega'_{-k_i,g}(\beta_{\mathbf{g}}T_i + \varepsilon_{ig})}{Np(1-p)}] \\
&= \frac{\sum_i E[\tilde{T}_i^2\omega'_{-k_i,g}\beta_{\mathbf{g}}]}{Np(1-p)} + \frac{\sum_i E[\tilde{T}_i\omega'_{-k_i,g}\varepsilon_{ig}]}{Np(1-p)} \\
&= \frac{\sum_i E[\tilde{T}_i^2]E[\omega'_{-k_i,g}\beta_{\mathbf{g}}]}{Np(1-p)} \\
&= \frac{p(1-p)\sum_i E[\omega'_{-k_i,g}\beta_{\mathbf{g}}]}{Np(1-p)} \\
&= E[\omega'_{-k_i,g}\beta_{\mathbf{g}}]
\end{aligned}$$

**Corollary 3.1.** *The OLS standard error for the $K$-fold optimus estimator $(\tilde{\mathbf{T}}'\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\tilde{\mathbf{Y}}_g$ is generally a biased estimate of the true standard error.*

Proof: Let $\varepsilon_g^\omega = [\omega'_{-1,g}\varepsilon_{1g}, \omega'_{-1,g}\varepsilon_{2g}, ..., \omega'_{-K,g}\varepsilon_{Ng}]'$. The OLS standard error for the $K$-fold optimus estimator relies on the assumption $E[\varepsilon_g^\omega\varepsilon_g^{\omega\prime}] = \sigma^2 I$. But $E[\omega'_{-k_i,g}\varepsilon_{ig} \cdot \omega'_{-k_j,g}\varepsilon_{jg}] \neq 0$, even when $i \neq j$, because $\omega_{-k_i,g}$ is a function of $\varepsilon_{-k_i,g}$, which generally includes $\varepsilon_{jg}$.

# A3 Shrinkage Estimator

In general the researcher does not know the covariance matrix for the outcomes, $\Sigma_g$, and must estimate it. While the sample covariance matrix $\hat{\Sigma}_g$ is consistent for $\Sigma_g$, in any finite sample the

off-diagonal elements suffer from a regression-to-the-mean problem: Large covariance values in the estimated matrix tend to be large both because the true $\sigma_{jk}$ (the covariance between outcomes $j$ and $k$) is non-zero and because there has been a stochastic shock in the sample correlation that is in the same direction as $\sigma_{jk}$. Using the raw estimate $\hat{\Sigma}_g$ thus tends to over (under) allocate weight to outcomes with abnormally high negative (positive) covariance entries. This suggests applying a shrinkage estimator to the off-diagonal elements of the estimated covariance matrix.[36]

The shrinkage estimator we consider is the Empirical Bayes estimator. This estimator applies Bayes Theorem:

$$\mathbb{P}(\sigma_{jk}|\hat{\sigma}_{jk}) = \frac{\mathbb{P}(\hat{\sigma}_{jk}|\sigma_{jk}) \cdot \mathbb{P}(\sigma_{jk})}{\mathbb{P}(\hat{\sigma}_{jk})}$$

Using the empirical distribution of the covariance entries for group $g$ and applying the law of iterated expectations we estimate:

$$\mathbb{P}(\sigma_{jk}|\hat{\sigma}_{jk}) = \frac{\mathbb{P}(\hat{\sigma}_{jk}|\sigma_{jk}) \cdot 2/(H_g^2 - H_g)}{\sum_{l=1}^{H_g} \sum_{m=l+1}^{H_g} \mathbb{P}(\hat{\sigma}_{jk}|\sigma_{jk} = \hat{\sigma}_{lm}) \cdot 2/(H_g^2 - H_g)} \tag{11}$$

In this context $H_g$ represents the number of outcomes in group $g$. Note that we only evaluate $\mathbb{P}(\sigma_{jk}|\hat{\sigma}_{jk})$ for values of $\sigma_{jk}$ corresponding to points of support in the empirical distribution of $\hat{\sigma}_{jk}$, and the denominator is a constant that ensures the posterior probabilities sum to one.[37] To understand the estimator's operation, consider the largest $\hat{\sigma}_{jk}$, $\hat{\sigma}_{max}$. The posterior for $\sigma_{max}$ is centered below $\hat{\sigma}_{max}$ because $\hat{\sigma}_{max}$ is the upper bound of the support for any posterior. Other coefficient estimates $\hat{\sigma}_{lm}$ "pull down" $\mathbb{E}[\sigma_{max}]$, with each posterior point of support $\hat{\sigma}_{lm}$ receiving weight $\mathbb{P}(\hat{\sigma}_{max}|\sigma_{max} = \hat{\sigma}_{lm})$. Thus the estimator "shrinks" larger covariance entries towards the empirical mean of the covariance entries. In practice we find that there tends to be too much shrinkage; for our final estimate of $\sigma_{jk}$ we use the (unweighted) average the Empirical Bayes estimate of $\sigma_{jk}$ (Equation (11)) and the raw estimate $\hat{\sigma}_{jk}$.

# A4 Analyses of Plans Incorporating Index Tests

This appendix analyzes the relative power of a rich variety of plans that combine index tests with exhaustive PAPs. First we establish that parallel test plans that include an index generally dominate equivalent plans that lack an index, as long as rejecting the index is of nontrivial value. Table A1 reports average power for a plan that tests an index, in parallel with other hypotheses, relative to the same plan without an index test. Panel A compares a PAP with an optimus index to the same PAP

[36] Note that the diagonal elements are all of similar magnitude due to the standardization of the outcomes.

[37] To compute $\mathbb{P}(\hat{\sigma}_{jk}|\sigma_{jk})$ we appeal to the Central Limit Theorem and assume an approximately normal distribution for $\hat{\sigma}_{jk}$.

without an optimus index, and Panel B compares a PAP with a KLK index to the same PAP without a KLK index. Due to the optimus index's power, plans with optimus indices are superior to their equivalents without indices even when rejecting the index is only half as valuable as rejecting an individual outcome (Panel A). PAPs with KLK indices are generally superior to their equivalents without indices when the index weight is 0.5 (Panel B), but the advantage is not as pronounced as it is with the optimus index (Panel A).

In high-powered cases in which many outcomes reject, rejecting the index along with many of the outcomes comprising the index may be of limited value. Table A2 reproduces Table A1 but applies a double-rejection adjustment so that the researcher does not receive double utility from rejecting an indicator by itself and as part of an index. To achieve this, the correction multiplies the index weight by $1 - a$, where $a$ is the fraction of index hypotheses that individually reject.[38] While the value of adding an index as an additional test falls with the double-rejection adjustment, the conclusions in Tables A1 and A2 are qualitatively similar. Thus, as long as researchers place a nontrivial weight on rejecting an index, it is advantageous to include the index test.

Next we analyze the tradeoffs between using an optimus index or a KLK index in a variety of scenarios. Table A3 presents the relative power of an optimus-gated PAP versus KLK index gated PAPs for different index weights. The table reveals how much more valuable the KLK index needs to be than the optimus index before researchers should switch from an optimus to a KLK index. The weight applied to the optimus index varies across rows, while the weight applied to the KLK index varies across columns. Panel A reports results for smaller families (Column (3) of Table A1), while Panel B reports results for larger families (Column (4) of Table A1). A weight of 1.0 implies that a researcher values an index rejection equivalently to rejecting a single outcome.

For small families, if the optimus index has a weight of 1, the KLK index weight needs to exceed 4 before a researcher is indifferent between using the KLK index or an optimus index.[39] For large families, if the optimus index has a weight of 1, the researcher prefers it to the KLK index even when she receives 8 times as much utility from rejecting the KLK index.

Table A4 presents, for different index weights, the relative power of PAPs that test an optimus index in parallel with other hypotheses versus those that test a KLK index in parallel. As above, these tables reveal how much more valuable the KLK index needs to be than the optimus index before researchers should switch from an optimus to a KLK index. Panel A reports results for smaller

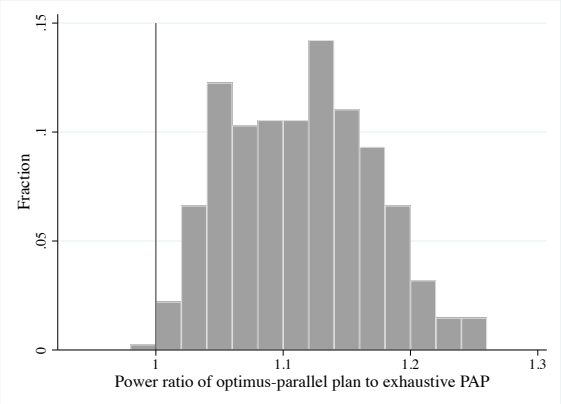
[38] Since $a$ will typically be much higher for the optimus index than the KLK index, the correction disproportionately affects the optimus.

[39] It is tempting to assume that the indifference point for an optimus weight of 2 should be a KLK index weight of 8. This logic ignores the dual role that the indices play, however — they generate rejections themselves, and they gate the testing of individual outcomes. Thus the power ratio of the two plans is not constant in the ratio of the two weights.

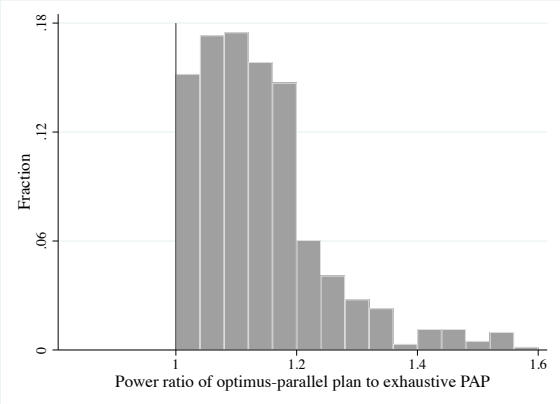
families (Column (3) of Table A1), while Panel B reports results for larger families (Column (4) of Table A1). For small families, if the optimus index has a weight of 1, the KLK index weight needs to exceed 5 before a researcher is indifferent between using the KLK index or an optimus index. For large families, if the optimus index has a weight of 1, the researcher prefers it to the KLK index even when she receives 6 times as much utility from rejecting the KLK index.

Figure A1: Distribution of Relative Power of Optimus Parallel Plan (with double-rejection adjustment) versus Exhaustive PAP

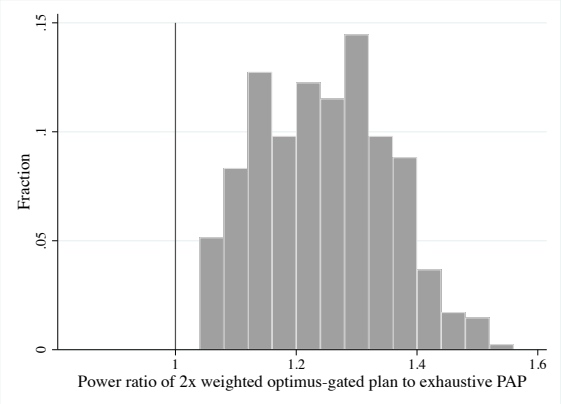
(a) Small families (index weight = 1)



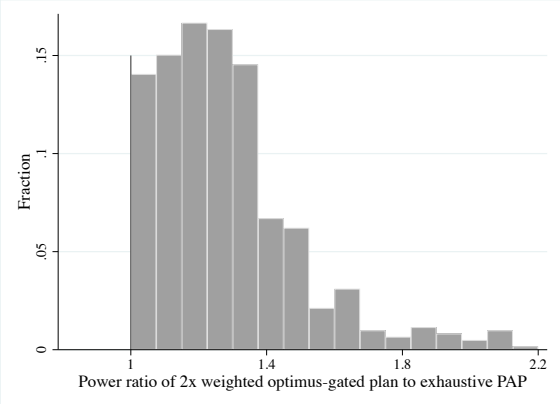

(b) Large families (index weight = 1)




(c) Small families (index weight = 2)




(d) Large families (index weight = 2)




*Notes:* Panels (a) and (c) correspond to Column (3) of Table 2, and Panels (b) and (d) correspond to Column (4). The double-rejection adjustment scales an index's rejection utility by one minus the fraction of index components that reject in the PAP.

Table A1: Relative Power of Plans w/ Indices v. Plans w/o Indices

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Index weight:* | A: Optimus + PAP v. PAP | | | | |
| 0.5 | 1.10 | 1.12 | 1.15 | 1.11 | 1.11 |
| 1.0 | 1.20 | 1.23 | 1.31 | 1.21 | 1.22 |
| 1.5 | 1.29 | 1.35 | 1.47 | 1.32 | 1.33 |
| 2.0 | 1.38 | 1.46 | 1.63 | 1.42 | 1.45 |
| | B: KLK index + PAP v. PAP | | | | |
| 0.5 | 1.04 | 1.04 | 1.01 | 1.02 | 1.00 |
| 1.0 | 1.09 | 1.08 | 1.04 | 1.04 | 1.02 |
| 1.5 | 1.13 | 1.12 | 1.08 | 1.06 | 1.03 |
| 2.0 | 1.17 | 1.15 | 1.11 | 1.08 | 1.05 |
| *Parameter restrictions:* | | | | | |
| Total hypotheses ($H$) | | | $\leq 20$ | $\geq 50$ | |
| Share false ($H_1/H$) | | $\leq 0.5$ | $\leq 0.2$ | $\leq 0.2$ | $0.1$ |
| Average effect size ($\mu_t$) | | $\leq 3.0$ | $\leq 3.0$ | $\leq 3.0$ | $\leq 2.5$ |
| Combinations | 2,600 | 1,560 | 416 | 624 | 390 |

Notes: Each cell reports, for a given index weight, the geometric mean power ratio of a plan that tests an index in parallel with other hypotheses to an equivalent plan that omits the index test. Panels A and B respectively test an optimus index with an exhaustive PAP and a KLK index with an exhaustive PAP.

Table A2: Relative Power of Plans w/ Indices v. Plans w/o Indices, Double-rejection Adjusted

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Index weight:* | A: Optimus + PAP v. PAP | | | | |
| 0.5 | 1.04 | 1.06 | 1.05 | 1.07 | 1.05 |
| 1.0 | 1.09 | 1.12 | 1.11 | 1.14 | 1.12 |
| 1.5 | 1.14 | 1.18 | 1.18 | 1.21 | 1.18 |
| 2.0 | 1.18 | 1.24 | 1.24 | 1.27 | 1.24 |
| | B: KLK index + PAP v. PAP | | | | |
| 0.5 | 1.03 | 1.03 | 1.01 | 1.02 | 1.00 |
| 1.0 | 1.07 | 1.07 | 1.04 | 1.04 | 1.02 |
| 1.5 | 1.11 | 1.10 | 1.06 | 1.06 | 1.03 |
| 2.0 | 1.14 | 1.14 | 1.09 | 1.07 | 1.04 |
| *Parameter restrictions:* | | | | | |
| Total hypotheses ($H$) | | | $\leq 20$ | $\geq 50$ | |
| Share false ($H_1/H$) | | $\leq 0.5$ | $\leq 0.2$ | $\leq 0.2$ | $0.1$ |
| Average effect size ($\mu_t$) | | $\leq 3.0$ | $\leq 3.0$ | $\leq 3.0$ | $\leq 2.5$ |
| Combinations | 2,600 | 1,560 | 416 | 624 | 390 |

Notes: Each cell reports, for a given index weight, the geometric mean power ratio of a plan that tests an index in parallel with other hypotheses to an equivalent plan that omits the index test. Panels A and B respectively test an optimus index with an exhaustive PAP and a KLK index with an exhaustive PAP. The double-rejection adjustment scales an index's rejection utility by one minus the fraction of index components that reject in the PAP.

Table A3: Relative Power of Optimus-Gated PAP v. KLK Index Gated PAP

| *KLK index weight:* | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 6.0 | 7.0 | 8.0 |
|---|---|---|---|---|---|---|---|---|
| *Optimus-index weight:* | | | | A: $H \leq 20$ | | | | |
| 1.0 | 2.41 | 1.66 | 1.27 | 1.03 | 0.87 | 0.75 | 0.66 | 0.59 |
| 2.0 | 3.54 | 2.44 | 1.87 | 1.52 | 1.28 | 1.10 | 0.97 | 0.87 |
| 3.0 | 4.67 | 3.22 | 2.47 | 2.00 | 1.68 | 1.45 | 1.28 | 1.14 |
| 4.0 | 5.79 | 4.00 | 3.06 | 2.48 | 2.09 | 1.80 | 1.59 | 1.42 |
| | | | | B: $H \geq 50$ | | | | |
| 1.0 | 2.54 | 2.06 | 1.74 | 1.52 | 1.35 | 1.21 | 1.10 | 1.01 |
| 2.0 | 3.27 | 2.65 | 2.24 | 1.95 | 1.73 | 1.56 | 1.42 | 1.30 |
| 3.0 | 3.97 | 3.22 | 2.72 | 2.37 | 2.11 | 1.89 | 1.72 | 1.58 |
| 4.0 | 4.66 | 3.77 | 3.20 | 2.78 | 2.47 | 2.22 | 2.02 | 1.86 |

Notes: Each cell reports, for a given combination of optimus-index and KLK index weights, the geometric mean power ratio of an optimus-index gated exhaustive PAP to a KLK index gated exhaustive PAP. Panel A uses the same set of parameter combinations as Column (3) of Table 2, and Panel B uses the same set of parameter combinations as Column (4) of Table 2 (416 and 624 parameter combinations respectively).

Table A4: Relative Power of Optimus-Parallel PAP v. KLK Index Parallel PAP

| *KLK index weight:* | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 6.0 | 7.0 | 8.0 |
|---|---|---|---|---|---|---|---|---|
| *Optimus-index weight:* | | | | A: $H \leq 20$ | | | | |
| 1.0 | 1.25 | 1.18 | 1.12 | 1.06 | 1.01 | 0.97 | 0.93 | 0.89 |
| 2.0 | 1.56 | 1.47 | 1.39 | 1.32 | 1.26 | 1.21 | 1.16 | 1.11 |
| 3.0 | 1.86 | 1.75 | 1.66 | 1.58 | 1.50 | 1.44 | 1.38 | 1.32 |
| 4.0 | 2.16 | 2.03 | 1.92 | 1.83 | 1.74 | 1.67 | 1.60 | 1.54 |
| | | | | B: $H \geq 50$ | | | | |
| 1.0 | 1.17 | 1.13 | 1.09 | 1.06 | 1.03 | 1.01 | 0.98 | 0.96 |
| 2.0 | 1.37 | 1.32 | 1.28 | 1.24 | 1.21 | 1.18 | 1.15 | 1.13 |
| 3.0 | 1.56 | 1.51 | 1.46 | 1.42 | 1.38 | 1.35 | 1.32 | 1.29 |
| 4.0 | 1.76 | 1.70 | 1.64 | 1.60 | 1.55 | 1.51 | 1.48 | 1.45 |

Notes: Each cell reports, for a given combination of optimus-index and KLK index weights, the geometric mean power ratio of an exhaustive PAP that tests the optimus index in parallel to an exhaustive PAP that tests the KLK index in parallel. Panel A uses the same set of parameter combinations as Column (3) of Table 2, and Panel B uses the same set of parameter combinations as Column (4) of Table 2 (416 and 624 parameter combinations respectively).

Table A5: Indicators and Weights in GoBiFo Optimus

| Hypothesis | Variable Name | Variable Label | Weight | PAP Family |
|---|---|---|---|---|
| Hardware | bank_acct | Does this community have a bank account? | 0.281 | H1,H3 |
| | vdc | Since January 2006, has this community had a Village or Community Development Co | 0.118 | H1 |
| | vdp | Does this community have a village development plan (i.e. an agreed plan with sp | 0.094 | H1 |
| | training | In the past 2 years (since October 2007), have you participated in any skills tr | 0.09 | H3 |
| | func_tba | Does the community have a traditional birth attendant (TBA) house and is it func | 0.089 | H2 |
| | f_comcntr | Does the community have a community center and is it functional? | 0.06 | H2 |
| | f_barrie | Does the community have a court barrie and is it functional? | 0.06 | H2 |
| | f_dryflr | Does the community have a drying floor and is it functional? | 0.05 | H2 |
| | att_wdc | Have you personally talked with a member of the WDC or participated in a meeting | 0.043 | H1 |
| | seedbank | Does this community have a seed bank (i.e. where people can borrow rice or groun | 0.028 | H2 |
| | quintile | Quintile of Household PCA Asset/Amenities score | 0.026 | H3 |
| | petty | [From supervisor tour of community] Have you seen anybody selling packaged goods | 0.021 | H3 |
| | tarp_public | [After asking the community how they have used (or plan to use) the tarp] SUPERV | 0.02 | H2 |
| | vis_wdc | Has this community been visited by a Ward Development Committee member in the pa | 0.015 | H1 |
| Software | commfarm | Does this community have any communal farms? | 0.091 | H4 |
| | no_fight | In the past 12 months, respondent has not been involved in any physical fighting | 0.076 | H11 |
| | minutes | Did anyone take minutes (written record of what was said) at the most recent com | 0.069 | H5 |
| | list_lc_chf | Relative view of 'do you think the Local Council [as opposed to Paramount chief] | 0.065 | H6 |
| | mbr_wom | Are you a member of any women's groups (general)? | 0.062 | H8 |
| | rtarp_public | [Given current chief chosen since 2005] Is the current (or acting) village chief | 0.06 | H6 |
| | name_elec | Correctly able to name the year of the next general elections | 0.055 | H9 |
| | leader_yth | Respondent agrees with 'Responsible young people can be good leaders' and not 'O | 0.054 | H6,H12 |
| | tstore_notchf | Village focus group says tarp is not stored in chief's private residence | 0.039 | H6 |
| | vote | Enumerator record of whether a vote occurred during the gift choice deliberation | 0.039 | H5,H6 |
| | name_sc | Correctly able to name the Section Chief for this section | 0.039 | H9 |
| | council_listen | Do you think the Local Council listens to what people in this town / neighborhoo | 0.037 | H10 |
| | trust_ngo | In your opinion, do you believe NGOs / donor projects or do you have to be caref | 0.035 | H7 |
| | mbr_seed | Are you a member of any seed multiplication groups? | 0.035 | H8 |
| | maj_gift | Gift (salt versus batteries) chosen reflects the view of the majority of househo | 0.03 | H5 |
| | mbr_trad | Are you a member of any traditional societies? | 0.029 | H8 |
| | say_tarp | Respondent feels that 'everybody in the village had equal say' in deciding what | 0.028 | H5 |
| | store_tarp | Tarp is stored in a public place (community center, school/clinic, church/mosque | 0.028 | H5 |
| | meet_yth | Enumerator record of total youths (18-35 years) present at gift choice meeting ( | 0.026 | H5 |
| | vis_pc | Has this community been visited by the Paramount Chief in the past year? | 0.019 | H9 |
| | bribebad | Respondent agrees with 'It's wrong to pay a bribe to any government official' an | 0.017 | H12 |
| | vh_fem | Is the current (or acting) village chief/Headman a woman? | 0.016 | H12 |
| | rmarket | Have you ever given money to a nonhousehold member to buy something for you at t | 0.014 | H7 |
| | dues | How much money have your given to church or mosque in the last month? [Add up al | 0.013 | H8 |
| | disabled_meet | Did any disabled people (blind, polio, amputee, wheelchair, etc.) attend the las | 0.012 | H5 |
| | notrad_cards | Respondent does not choose a chiefdom official or elder in response to 'who had | 0.01 | H6 |

| Hypothesis | Variable Name | Variable Label | Weight | PAP Family |
|---|---|---|---|---|
| Software (no community farm) | no_fight | In the past 12 months, respondent has not been involved in any physical fighting | 0.086 | H11 |
| | minutes | Did anyone take minutes (written record of what was said) at the most recent com | 0.073 | H5 |
| | mbr_wom | Are you a member of any women's groups (general)? | 0.071 | H8 |
| | list_lc_chf | Relative view of 'do you think the Local Council [as opposed to Paramount chief] | 0.066 | H6 |
| | rtarp_public | [Given current chief chosen since 2005] Is the current (or acting) village chief | 0.062 | H6 |
| | name_elec | Correctly able to name the year of the next general elections | 0.06 | H9 |
| | leader_yth | Respondent agrees with 'Responsible young people can be good leaders' and not 'O | 0.06 | H6,H12 |
| | council_listen | Do you think the Local Council listens to what people in this town / neighborhoo | 0.049 | H10 |
| | tstore_notchf | Village focus group says tarp is not stored in chief's private residence | 0.046 | H6 |
| | trust_ngo | In your opinion, do you believe NGOs / donor projects or do you have to be caref | 0.041 | H7 |
| | mbr_seed | Are you a member of any seed multiplication groups? | 0.039 | H8 |
| | vote | Enumerator record of whether a vote occurred during the gift choice deliberation | 0.039 | H5,H6 |
| | name_sc | Correctly able to name the Section Chief for this section | 0.038 | H9 |
| | meet_yth | Enumerator record of total youths (18-35 years) present at gift choice meeting ( | 0.032 | H5 |
| | say_tarp | Respondent feels that 'everybody in the village had equal say' in deciding what | 0.031 | H5 |
| | maj_gift | Gift (salt versus batteries) chosen reflects the view of the majority of househo | 0.03 | H5 |
| | store_tarp | Tarp is stored in a public place (community center, school/clinic, church/mosque | 0.029 | H5 |
| | mbr_trad | Are you a member of any traditional societies? | 0.026 | H8 |
| | vis_pc | Has this community been visited by the Paramount Chief in the past year? | 0.021 | H9 |
| | bribebad | Respondent agrees with 'It's wrong to pay a bribe to any government official' an | 0.019 | H12 |
| | vh_fem | Is the current (or acting) village chief/Headman a woman? | 0.016 | H12 |
| | rmarket | Have you ever given money to a nonhousehold member to buy something for you at t | 0.016 | H7 |
| | dues | How much money have your given to church or mosque in the last month? [Add up al | 0.014 | H8 |
| | disabled_meet | Did any disabled people (blind, polio, amputee, wheelchair, etc.) attend the las | 0.012 | H5 |
| | duration | Enumerator record of duration of gift choice deliberation in minutes (field acti | 0.011 | H5 |
| | notrad_cards | Respondent does not choose a chiefdom official or elder in response to 'who had | 0.01 | H6 |
| | spend_lc_chf | Relative view of 'if the Local Council [as opposed to Paramount chief] was given | 0.01 | H6 |
| Hypothesis 1 | bank_acct | Does this community have a bank account? | 0.5 | H1,H3 |
| | vdc | Since January 2006, has this community had a Village or Community Development Co | 0.227 | H1 |
| | vdp | Does this community have a village development plan (i.e. an agreed plan with sp | 0.144 | H1 |
| | att_wdc | Have you personally talked with a member of the WDC or participated in a meeting | 0.082 | H1 |
| | vis_wdc | Has this community been visited by a Ward Development Committee member in the pa | 0.05 | H1 |
| Hypothesis 2 | func_tba | Does the community have a traditional birth attendant (TBA) house and is it func | 0.201 | H2 |
| | seedbank | Does this community have a seed bank (i.e. where people can borrow rice or groun | 0.177 | H2 |
| | f_barrie | Does the community have a court barrie and is it functional? | 0.172 | H2 |
| | f_comcntr | Does the community have a community center and is it functional? | 0.153 | H2 |
| | f_latrine | Does the community have a latrine and is it functional? | 0.113 | H2 |
| | f_dryflr | Does the community have a drying floor and is it functional? | 0.077 | H2 |
| | footunif | Do any of the local sports teams have uniforms / vests? | 0.046 | H2 |
| | tarp_public | [After asking the community how they have used (or plan to use) the tarp] SUPERV | 0.03 | H2 |
| | func_sports | Does the community have a football / sports field and is it functional? | 0.021 | H2 |
| | f_psch | Does the community have a primary school and is it functional? | 0.021 | H2 |

| Hypothesis | Variable Name | Variable Label | Weight | PAP Family |
|---|---|---|---|---|
| Hypothesis 3 | bank_acct | Does this community have a bank account? | 0.6 | H1,H3 |
| | training | In the past 2 years (since October 2007), have you participated in any skills tr | 0.186 | H3 |
| | quintile | Quintile of Household PCA Asset/Amenities score | 0.071 | H3 |
| | assets | Household PCA Asset/Amenities score (includes hhs ownership of bicycle, mobile p | 0.039 | H3 |
| | tot_goods | Number of goods out of 10 common items (bread, soap, garri, country cloth/garra | 0.036 | H3 |
| | petty | [From supervisor tour of community] Have you seen anybody selling packaged goods | 0.028 | H3 |
| | betteroff | Supervisor assessment that community is 'much better off' or 'a little better of | 0.021 | H3 |
| | tot_petty | How many houses and small shops (including tables, boxes and kiosks) are selling | 0.02 | H3 |
| Hypothesis 4 | commfarm | Does this community have any communal farms? | 0.713 | H4 |
| | vchr_tot | How much money do you think the community will be able to raise to use the build | 0.13 | H4 |
| | mkt_grp | Do any people from different households here come together to sell agricultural | 0.054 | H4,H7,H8 |
| | cards | Number of vouchers for building materials out of 6 maximum that the community re | 0.046 | H4 |
| | wkcomfrm | In the past one year, did you work on a communal farm (this means a farm owned b | 0.041 | H4 |
| Hypothesis 5 | minutes | Did anyone take minutes (written record of what was said) at the most recent com | 0.014 | H5 |
| | say_tarp | Respondent feels that 'everybody in the village had equal say' in deciding what | 0.01 | H5 |
| | show_tarp | Supervisor asks to see the tarp at second round follow-up visit: can the communi | 0.01 | H5 |
| Hypothesis 6 | rtarp_public | [Given current chief chosen since 2005] Is the current (or acting) village chief | 0.218 | H6 |
| | list_lc_chf | Relative view of 'do you think the Local Council [as opposed to Paramount chief] | 0.177 | H6 |
| | leader_yth | Respondent agrees with 'Responsible young people can be good leaders' and not 'O | 0.16 | H6,H12 |
| | tstore_notchf | Village focus group says tarp is not stored in chief's private residence | 0.157 | H6 |
| | notrad_tarp | Respondent does not choose a chiefdom official or elder in response to 'who had | 0.113 | H6 |
| | vote | Enumerator record of whether a vote occurred during the gift choice deliberation | 0.091 | H5,H6 |
| | notrad_cards | Respondent does not choose a chiefdom official or elder in response to 'who had | 0.028 | H6 |
| | leader_wmn | Respondent agrees with 'Women can be good politicians and should be encouraged t | 0.028 | H6,H12 |
| | spend_lc_chf | Relative view of 'if the Local Council [as opposed to Paramount chief] was given | 0.017 | H6 |
| | question_auth | Respondent agrees with 'As citizens, we should be more active in questioning the | 0.016 | H6 |
| Hypothesis 7 | trust_ngo | In your opinion, do you believe NGOs / donor projects or do you have to be caref | 0.471 | H7 |
| | rmarket | Have you ever given money to a nonhousehold member to buy something for you at t | 0.189 | H7 |
| | hmarket | Tomorrow, if you needed to buy something from town or the market but were unable | 0.062 | H7 |
| | osusu | Are you a member of any credit or savings (osusu) groups? | 0.032 | H7,H8 |
| | trust_pol | In your opinion, do you believe the police or do you have to be careful when dea | 0.025 | H7 |
| | trust_own | In your opinion, do you believe people from you own village / town / neighborhoo | 0.014 | H7 |
| | trust_out | In your opinion, do you believe people from outside you own village / town / nei | 0.012 | H7 |

| Hypothesis | Variable Name | Variable Label | Weight | PAP Family |
|---|---|---|---|---|
| Hypothesis 8 | mbr_wom | Are you a member of any women's groups (general)? | 0.524 | H8 |
| | mbr_seed | Are you a member of any seed multiplication groups? | 0.259 | H8 |
| | mbr_trad | Are you a member of any traditional societies? | 0.118 | H8 |
| | dues | How much money have your given to church or mosque in the last month? [Add up al | 0.055 | H8 |
| | osusu | Are you a member of any credit or savings (osusu) groups? | 0.021 | H7,H8 |
| Hypothesis 9 | name_elec | Correctly able to name the year of the next general elections | 0.377 | H9 |
| | name_sc | Correctly able to name the Section Chief for this section | 0.227 | H9 |
| | name_chr | Correctly able to name the Chairperson of the Local Council | 0.125 | H9 |
| | vis_pc | Has this community been visited by the Paramount Chief in the past year? | 0.1 | H9 |
| | disp_ind | Supervisor assessment of whether there are any of the following items–awareness | 0.034 | H9 |
| | radio | Do you get information from the radio about politics and what the government is | 0.027 | H9 |
| Hypothesis 10 | vote_local | Did you vote in the local government election (2008)? | 0.206 | H10 |
| | stand_lc | Did anyone in this community contest the party symbol in the 2008 local council | 0.039 | H10 |
| | change_council | Respondent thinks they have 'some' or 'little' as opposed to 'no' chance to chan | 0.029 | H10 |
| | vote_pres1 | Enumerator verifies that respondent's voter ID card has the correct hole punched | 0.029 | H10 |
| | cvote_local | Enumerator verifies that respondent's voter ID card has the correct hole punched | 0.022 | H10 |
| Hypothesis 11 | no_fight | In the past 12 months, respondent has not been involved in any physical fighting | 0.599 | H11 |
| | no_conflict | No conflict that respondent needed help from someone outside the household to re | 0.23 | H11 |
| | nobeatchild | Respondent agrees with 'Beating children will only teach them to use violence ag | 0.035 | H11 |
| | no_witch | During the last 12 months, respondent has not been a victim of witchcraft (juju) | 0.03 | H11 |
| | violence_bad | Respondent agrees with 'The use of violence is never justified in politics' and | 0.028 | H11 |
| | no_theft | In the past 12 months, no livestock, household items or money stolen from the re | 0.013 | H11 |
| Hypothesis 12 | leader_yth | Respondent agrees with 'Responsible young people can be good leaders' and not 'O | 0.318 | H6,H12 |
| | bribebad | Respondent agrees with 'It's wrong to pay a bribe to any government official' an | 0.164 | H12 |
| | vh_fem | Is the current (or acting) village chief/Headman a woman? | 0.135 | H12 |
| | leader_wmn | Respondent agrees with 'Women can be good politicians and should be encouraged t | 0.104 | H6,H12 |
| | youthtreat | Respondent agress with 'In this community, elders / authorities treat youths jus | 0.016 | H12 |

Table A6: Disaggregated OHIE Results (100% Sample)

| Indicator (standardized outcome) | ITT estimate | RW $p$-val |
|---|---|---|
| *Subfamily: Utilization, prevention, access* | | |
| Currently taking any prescription medications | 0.039 | 0.845 |
| | (0.023) | |
| Any primary care visits | 0.103 | 0.000 |
| | (0.023) | |
| Any ER visits | 0.021 | 1.000 |
| | (0.023) | |
| Any hospital visits | -0.009 | 1.000 |
| | (0.023) | |
| Number of prescription meds currently taking | 0.069 | 0.090 |
| | (0.023) | |
| Number of primary care visits | 0.098 | 0.000 |
| | (0.023) | |
| Number of ER visits | 0.012 | 1.000 |
| | (0.023) | |
| Number of hospital visits | 0.012 | 1.000 |
| | (0.023) | |
| Ever had cholesterol checked | 0.031 | 0.965 |
| | (0.024) | |
| Ever had diabetes checked | 0.059 | 0.245 |
| | (0.023) | |
| Ever had a mammogram | 0.067 | 0.105 |
| | (0.023) | |
| Ever had a pap smear | 0.066 | 0.105 |
| | (0.022) | |
| Ever had diabetes/sugar diabetes diagnosis | -0.003 | 1.000 |
| | (0.023) | |
| Ever had asthma diagnosis | -0.013 | 1.000 |
| | (0.023) | |
| Ever had high blood pressure diagnosis | 0.014 | 1.000 |
| | (0.023) | |

| Indicator (standardized outcome) | (1) ITT estimate | (2) RW $p$-val |
|---|---|---|
| Ever had COPD diagnosis | 0.047 | 0.590 |
| | (0.023) | |
| Ever had heart disease/angina diagnosis | -0.036 | 0.890 |
| | (0.023) | |
| Ever had congestive heart failure diagnosis | 0.016 | 1.000 |
| | (0.024) | |
| Ever had depression/anxiety diagnosis | -0.007 | 1.000 |
| | (0.023) | |
| Ever had high cholesterol diagnosis | 0.011 | 1.000 |
| | (0.023) | |
| Ever had kidney disease diagnosis | -0.027 | 0.975 |
| | (0.023) | |
| Usual place of care is clinic | 0.154 | 0.000 |
| | (0.024) | |
| Have personal doctor | 0.129 | 0.000 |
| | (0.024) | |
| Got all needed medical care in last 6 months | 0.158 | 0.000 |
| | (0.023) | |
| Got all needed prescriptions in last 6 months | 0.105 | 0.000 |
| | (0.023) | |
| Did not use ER for non-ER care | -0.002 | 1.000 |
| | (0.023) | |
| *Subfamily: Health* | | |
| Overall health excellent/good | 0.051 | 0.490 |
| | (0.023) | |
| Overall health not poor | 0.047 | 0.620 |
| | (0.023) | |
| Change in overall health (positive is better) | 0.097 | 0.000 |
| | (0.023) | |
| Number of days (in past 30) not impaired by poor health | 0.013 | 1.000 |
| | (0.023) | |
| Number of days (in past 30) when physical health good | 0.009 | 1.000 |
| | (0.023) | |

| Indicator (standardized outcome) | (1) ITT estimate | (2) RW $p$-val |
|---|---|---|
| Number of days (in past 30) when mental health good | -0.002 | 1.000 |
| | (0.024) | |
| Not depressed in past 2 weeks | -0.005 | 1.000 |
| | (0.023) | |
| Not current smoker | -0.029 | 0.965 |
| | (0.023) | |
| Physical activity (compared to others of same age) | -0.027 | 0.975 |
| | (0.023) | |
| Current overall happiness (higher is better) | 0.063 | 0.150 |
| | (0.023) | |
| *Subfamily: Financial* | | |
| Household income as percent of federal poverty line | 0.009 | 1.000 |
| | (0.024) | |
| Household income category | 0.016 | 1.000 |
| | (0.024) | |
| Currently employed | -0.003 | 1.000 |
| | (0.023) | |
| Average weekly hours worked | 0 | 1.000 |
| | (0.023) | |
| No out of pocket costs for medical care in past 6 months | 0.183 | 0.000 |
| | (0.023) | |
| Do not currently owe money for medical expenses | 0.045 | 0.665 |
| | (0.023) | |
| Haven't borrowed to pay health care bills in past 6 months | 0.109 | 0.000 |
| | (0.024) | |
| Haven't been refused care because owed money for past treatment | 0.048 | 0.620 |
| | (0.023) | |

Notes: Results in Column (1) represent coefficients from a regression of the listed indicator (standardized to unit variance) on an intention-to-treat indicator using the 100% sample ($N = 8,141$), controlling for household size and survey round fixed effects. Parentheses contain standard errors clustered at the household level. Column (2) reports Romano-Wolf $p$-values that control FWER across the 44 outcomes in the table.

Table A7: Indicators and Weights in OHIE

| Hypothesis | Variable Name | Variable Label | Weight |
|---|---|---|---|
| All Outcomes | neg_cost_any_oop_12m | No out of pocket costs for medical care in past 6 months | 0.194 |
| | usual_clinic_12m | Usual place of care is clinic | 0.088 |
| | doc_any_12m | Any primary care visits | 0.067 |
| | needmet_med_12m | Got all needed medical care in last 6 months | 0.060 |
| | needmet_rx_12m | Got all needed prescriptions in last 6 months | 0.044 |
| | health_chgflip_bin_12m | Change in overall health (positive is better) | 0.040 |
| | usual_doc_12m | Have personal doctor | 0.036 |
| | neg_cost_borrow_12m | Haven't borrowed to pay health care bills in past 6 months | 0.033 |
| | doc_num_mod_12m | Number of primary care visits | 0.033 |
| | emp_dx_12m | Ever had COPD diagnosis | 0.032 |
| | rx_num_mod_12m | Number of prescription meds currently taking | 0.027 |
| | dia_chk_bin_12m | Ever had diabetes checked | 0.026 |
| | mam_chk_bin_all_12m | Ever had a mammogram | 0.026 |
| | rx_any_12m | Currently taking any prescription medications | 0.023 |
| | pap_chk_bin_all_12m | Ever had a pap smear | 0.023 |
| | neg_cost_refused_12m | Haven't been refused care because owed money for past treatment | 0.017 |
| | neg_cost_any_owe_12m | Do not currently owe money for medical expenses | 0.016 |
| | poshappiness_bin_12m | Current overall happiness (higher is better) | 0.016 |
| | health_notpoor_12m | Overall health not poor | 0.016 |
| | chl_dx_12m | Ever had high cholesterol diagnosis | 0.013 |
| | hhinc_pctfpl_12m | Household income as percent of federal poverty line | 0.012 |
| | chf_dx_12m | Ever had congestive heart failure diagnosis | 0.012 |
| | er_any_12m | Any ER visits | 0.012 |
| | er_num_mod_12m | Number of ER visits | 0.012 |
| | health_genflip_bin_12m | Overall health excellent/good | 0.012 |
| | hbp_dx_12m | Ever had high blood pressure diagnosis | 0.010 |
| | chl_chk_bin_12m | Ever had cholesterol checked | 0.010 |
| | employ_hrs_12m | Average weekly hours worked | 0.010 |
| | hhinc_cat_12m | Household income category | 0.010 |

| Hypothesis | Variable Name | Variable Label | Weight |
|---|---|---|---|
| | needmet_med_12m | Got all needed medical care in last 6 months | 0.177 |
| | usual_clinic_12m | Usual place of care is clinic | 0.129 |
| | needmet_rx_12m | Got all needed prescriptions in last 6 months | 0.123 |
| | doc_any_12m | Any primary care visits | 0.052 |
| | emp_dx_12m | Ever had COPD diagnosis | 0.050 |
| | usual_doc_12m | Have personal doctor | 0.050 |
| | doc_num_mod_12m | Number of primary care visits | 0.045 |
| | dia_chk_bin_12m | Ever had diabetes checked | 0.039 |
| | pap_chk_bin_all_12m | Ever had a pap smear | 0.038 |
| | mam_chk_bin_all_12m | Ever had a mammogram | 0.037 |
| Utilization, | rx_num_mod_12m | Number of prescription meds currently taking | 0.026 |
| prevention, | chf_dx_12m | Ever had congestive heart failure diagnosis | 0.019 |
| access | er_any_12m | Any ER visits | 0.018 |
| | chl_dx_12m | Ever had high cholesterol diagnosis | 0.017 |
| | not_er_noner_12m | Did not use ER for non-ER care | 0.016 |
| | hosp_num_mod_12m | Number of hospital visits | 0.015 |
| | hbp_dx_12m | Ever had high blood pressure diagnosis | 0.014 |
| | chl_chk_bin_12m | Ever had cholesterol checked | 0.014 |
| | rx_any_12m | Currently taking any prescription medications | 0.014 |
| | dep_dx_12m | Ever had depression/anxiety diagnosis | 0.014 |
| | er_num_mod_12m | Number of ER visits | 0.013 |
| | dia_dx_12m | Ever had diabetes/sugar diabetes diagnosis | 0.011 |
| | ast_dx_12m | Ever had asthma diagnosis | 0.011 |

| Hypothesis | Variable Name | Variable Label | Weight |
|---|---|---|---|
| Health | health_chgflip_bin_12m | Change in overall health (positive is better) | 0.198 |
| | poshappiness_bin_12m | Current overall happiness (higher is better) | 0.123 |
| | nonsmk_curr_12m | Not current smoker | 0.097 |
| | health_notpoor_12m | Overall health not poor | 0.094 |
| | more_active_12m | Physical activity (compared to others of same age) | 0.072 |
| | health_genflip_bin_12m | Overall health excellent/good | 0.062 |
| | nodep_screen_12m | Not depressed in past 2 weeks | 0.052 |
| | notbaddays_tot_12m | Number of days (in past 30) not impaired by poor health | 0.038 |
| | notbaddays_phys_12m | Number of days (in past 30) when physical health good | 0.032 |
| | notbaddays_ment_12m | Number of days (in past 30) when mental health good | 0.026 |
| Financial | neg_cost_any_oop_12m | No out of pocket costs for medical care in past 6 months | 0.399 |
| | neg_cost_borrow_12m | Haven't borrowed to pay health care bills in past 6 months | 0.142 |
| | neg_cost_refused_12m | Haven't been refused care because owed money for past treatment | 0.110 |
| | hhinc_pctfpl_12m | Household income as percent of federal poverty line | 0.072 |
| | neg_cost_any_owe_12m | Do not currently owe money for medical expenses | 0.066 |
| | hhinc_cat_12m | Household income category | 0.062 |
| | employ_hrs_12m | Average weekly hours worked | 0.045 |
| | employ_12m | Currently employed | 0.038 |

Notes: Hypothesis refers to the family or subfamily of indicators used to construct the optimus index. Weight represents average weight received by an indicator variable across 100 random 10% samples.

Table A8: OHIE Results by Sample Size: Gating Hypotheses and Individual Indicators

| Test | (1) Power | (2) Effect Size | (3) Index Size | (4) True Effect Size |
|---|---|---|---|---|
| | Panel A: 8% sample | | | |
| Optimus | 54% | 0.086 | 8.8 | 0.092 |
| KLK Index | 50% | 0.041 | 44 | 0.039 |
| LMS Omnibus | 38% | | | |
| Exhaustive PAP | 48% | $0.211^{+}$ | $1.3^{++}$ | 0.115 |
| PAP (post optimus gate) | 34% | $0.211^{+}$ | $1.2^{++}$ | 0.115 |
| | Panel B: 12% sample | | | |
| Optimus | 85% | 0.097 | 9.6 | 0.102 |
| KLK Index | 71% | 0.04 | 44 | 0.039 |
| LMS Omnibus | 57% | | | |
| Exhaustive PAP | 67% | $0.207^{+}$ | $1.6^{++}$ | 0.136 |
| PAP (post optimus gate) | 61% | $0.207^{+}$ | $1.5^{++}$ | 0.137 |
| | Panel C: 15% sample | | | |
| Optimus | 90% | 0.097 | 9.9 | 0.105 |
| KLK Index | 77% | 0.037 | 44 | 0.039 |
| LMS Omnibus | 57% | | | |
| Exhaustive PAP | 75% | $0.209^{+}$ | $1.8^{++}$ | 0.141 |
| PAP (post optimus gate) | 70% | $0.209^{+}$ | $1.8^{++}$ | 0.141 |

Notes: Results in Columns (1) – (3) represent averages across 100 random 8%, 12%, or 15% samples of the OHIE data. Power denotes power to reject the sharp null hypothesis for at least one indicator. True effect size represents the estimated effect in the full (100% sample) OHIE dataset, with indicators weighted using average weights underlying Column (2).

+ Average effect size for indicators rejected by the PAP (when no indicator rejects, calculation includes the most significant indicator).

++ Average number of indicators rejected, left-censored at 1.

Table A9: OHIE Results: Ungated Subfamily Hypotheses

| Test | (1) Power | (2) Effect Size | (3) Index Size | (4) True Effect Size |
|---|---|---|---|---|
| | Panel A: Utilization outcomes | | | |
| Optimus | 36% | 0.087 | 5.6 | 0.094 |
| KLK Index | 21% | 0.042 | 26 | 0.043 |
| | Panel B: Health outcomes | | | |
| Optimus | 7% | 0.028 | 2.3 | 0.038 |
| KLK Index | 2% | 0.021 | 10 | 0.022 |
| | Panel C: Financial outcomes | | | |
| Optimus | 23% | 0.095 | 2.7 | 0.105 |
| KLK Index | 17% | 0.049 | 8 | 0.051 |

Notes: Results in Columns (1) – (3) represent averages across 100 random 10% samples of the OHIE data. Power denotes power to reject the sharp null hypothesis for at least one subfamily indicator, multiplicity adjusted for the three parallel subfamily tests. True effect size represents the estimated effect in the full (100% sample) OHIE dataset, with indicators weighted using average weights underlying Column (2).

Table A10: OHIE Results: Gated Subfamily Hypotheses

| Test | (1) Power | (2) Effect Size | (3) Index Size | (4) True Effect Size |
|---|---|---|---|---|
| | Panel A: Utilization outcomes | | | |
| Optimus | 34% | 0.087 | 5.6 | 0.094 |
| KLK Index | 20% | 0.042 | 26 | 0.043 |
| | Panel B: Health outcomes | | | |
| Optimus | 7% | 0.028 | 2.3 | 0.038 |
| KLK Index | 2% | 0.021 | 10 | 0.022 |
| | Panel C: Financial outcomes | | | |
| Optimus | 22% | 0.095 | 2.7 | 0.105 |
| KLK Index | 15% | 0.049 | 8 | 0.051 |

Notes: Results in Columns (1) – (3) represent averages across 100 random 10% samples of the OHIE data. Power denotes power to pass the all-outcome gate and reject the sharp null hypothesis for at least one subfamily indicator, multiplicity adjusted for the three parallel subfamily tests. True effect size represents the estimated effect in the full (100% sample) OHIE dataset, with indicators weighted using average weights underlying Column (2).

Table A11: OHIE Results: Ungated OHIE PAP Table Hypotheses

| Test | (1) Power | (2) Effect Size | (3) Index Size |
|---|---|---|---|
| | Panel A: Tables P1 + U1 | | |
| Optimus | 13% | 0.055 | 2.2 |
| KLK Index | 9% | 0.038 | 8 |
| | Panel B: Table P2 | | |
| Optimus | 57% | -0.127 | 1.7 |
| KLK Index | 44% | -0.097 | 4 |
| | Panel C: Table P3 | | |
| Optimus | 23% | 0.050 | 1.9 |
| KLK Index | 8% | 0.034 | 7 |

Notes: Results in Columns (1) – (3) represent averages across 100 random 10% samples of the OHIE data. Power denotes power to reject the sharp null hypothesis for at least one subfamily indicator, with no multiplicity adjustment. Each panel includes outcomes from the referenced tables in Finkelstein et al. (2010). Tables P1 and U1 contain eight utilization outcomes, Table P2 contains four financial strain outcomes, and Table P3 contains seven health outcomes.