Optimal Pre-Analysis Plans: Statistical Decisions Subject to Implementability

(preliminary draft)

Maximilian Kasy Jann Spiess

October 6, 2023

Abstract

What is the purpose of pre-analysis plans, and how should they be designed? We propose a principal–agent model where a decision-maker relies on selective but truthful reports by an analyst. The analyst has data access, and non-aligned objectives. In this model, the implementation of statistical decision rules (tests, estimators) requires an incentive-compatible mechanism. We first characterize which decision rules can be implemented. We then characterize optimal statistical decision rules subject to implementability. We show that implementation requires pre-analysis plans. Focussing specifically on hypothesis tests, we show that optimal rejection rules pre-register a valid test for the case when all data is reported, and make worst-case assumptions about unreported data. Optimal tests can be found as a solution to a linear-programming problem.

Keywords: Pre-analysis plans, Statistical decisions, Implementability
JEL codes: C18, D8, I23

We thank Alex Frankel, Carlos Gonzalez Perez, Rohit Lamba, Pascal Michaillat, Ted Miguel, Marco Ottaviani, Ludvig Sinander, Alex Teytelboym, and participants at the BITSS 2022 meeting, the 2022 AEA meetings, and the 2022 conference in Honor of Jim Powell for helpful discussions and suggestions.
Maximilian Kasy was supported by the Alfred P. Sloan Foundation, under the grant "Social foundations for statistics and machine learning."
A precursor manuscript with a different model and results on implementability with costly communication, titled "Rationalizing Pre-Analysis Plans: Statistical Decisions Subject to Implementability", is available at https://arxiv.org/abs/2208.09638v1.

1 Introduction

When writing up their studies, empirical researchers might cherry-pick the findings that they report. Cherry-picking distorts the inferences that we can draw from published findings, cf. Andrews and Kasy (2019); Andrews et al. (2023). As a potential remedy, pre-analysis plans (PAPs) have become a precondition for the publication of experimental research in economics, for both field experiments and lab experiments.[1] PAPs might enable valid inference by pre-specifying a mapping from the data to testing decisions or estimates, cf. Christensen and Miguel (2018); Miguel (2021). This might prevent the cherry-picking of results, and thus provide a remedy for the distortions introduced by unacknowledged multiple hypothesis testing. The widespread adoption of PAPs has not gone uncontested, however,[2] and has been criticized for constraining our ability to learn from experiments.

In this article, we aim to clarify the benefits and optimal design of pre-analysis plans by modeling statistical inference as a mechanism-design problem (Myerson, 1986; Kamenica, 2019; Sinander, 2023). To motivate this approach, note that, in single-agent statistical decision theory, rational decision-makers with preferences that are consistent over time have no need for the commitment device that is provided by a PAP. This holds in particular when a single decision-maker aims to construct tests that control size, or estimators that are unbiased; they have no reason to "cheat themselves." The situation is different, however, when there are multiple agents with conflicting interests, or when preferences are not consistent over time.

In this article, we provide detailed guidance for practitioners, including both decision-makers (e.g., readers, editors) and data analysts (e.g., study authors). From the decision-makers' perspective, we describe how tests, estimators, or other decision rules can be implemented by requiring pre-analysis plans. We then focus on hypothesis tests in particular, and show, from the analysts' perspective, how to derive optimal pre-analysis plans, which maximize power while controlling size and maintaining implementability. We also provide software (an interactive web app) to facilitate the design of optimal pre-analysis plans.

[1] Just as in the case of randomized experiments, the adoption of PAPs in economics follows their prior adoption in clinical research; see for instance the guidelines of the FDA on PAPs, (Food and Drug Administration, 1998).

[2] See for instance Coffman and Niederle (2015), Olken (2015), and Duflo et al. (2020), who discuss the costs and benefits of PAPs in experimental economics from a practitioners' perspective.

Examples In our model, we consider the interaction between a decision-maker and an analyst with private information and conflicting interests. One example of such a conflict of interest is between a researcher (analyst) who wants to reject a hypothesis, and a reader of their research (decision-maker) who wants a valid statistical test of that same hypothesis; the relevant decision here is whether to reject the null hypothesis. Another example is the conflict of interest between a researcher (analyst) who wants to get published, and a journal editor (decision-maker) who only wants to publish studies on effects that are large enough to be interesting; the relevant decision here is whether to publish a study. A third example is the conflict of interest between a pharmaceutical company (analyst) who wants to sell drugs, and a medical regulatory agency (decision-maker) who wants to protect patient health; the relevant decision here is whether to approve a drug.

Model and timeline The mechanism-design approach that we propose takes the perspective of a decision-maker who wants to implement a statistical decision rule. Not all rules are implementable, however, when the analyst has divergent interests and private information. This mechanism-design perspective allows us to stay close to standard statistical theory, while taking into account the constraints that are a consequence of the social nature of research.

The timeline of our model is as follows. Before observing the data, the analyst might send a pre-analysis message M to the decision-maker. This message might be in the form of a pre-analysis plan. The full data which are potentially observed by the analyst are given by a vector $X = (X_1, \ldots, X_k)$, and are distributed according to some true, unknown state θ. The components of X might, for instance, correspond to the outcomes of different hypothesis tests, or to estimates for different model specifications. Not all of these components are actually available to the analyst, however. Instead, the analyst only observes X_i for $i \in J$, where the set of available components $J \subseteq \{1, \ldots, k\}$ is private information of the analyst. The analyst gets to choose a subset $I \subseteq J$ of observations, and then reports (X_I, I) to the decision-maker, who takes a decision $A \in \mathcal{A} \subseteq \mathbb{R}$. The analyst always prefers a higher value for A. We consider different objectives for the decision-maker, including, in particular, statistical testing subject to size control, in which case A is the probability of rejecting the null hypothesis.

In this model, the analyst might *hide* information from the decision-maker, but they may not *lie* about the components they report. Put differently, our model is one of partial verifiability. The potential value of a pre-analysis message in this setting comes from the fact that it allows the analyst to share private information (i.e., expertise) with the decision-maker, in a way that would not be incentive compatible if a message could only be sent after seeing the data. The analyst might have private information regarding the availability of components J, and regarding the state of the world θ. Availability of components X_i might be restricted for various reasons: Not every estimate or test that is a priori conceivable is actually at the analyst's disposition. Prior uncertainty about component availability allows for plausible deniability by the analyst. Experiments might not have been run, or data might not have been collected.

Implementable decision rules We first characterize the set of implementable statistical decision rules, which is independent of decision-maker preferences. We show that implementable decision rules are required to be *monotonic in the reported set* of components I, in terms of set inclusion, given the pre-analysis message, and given X. Put differently, reporting more results can never make the analyst worse off. Implementable decision rules furthermore need to be compatible with *truthful revelation of analyst private information* prior to observing any data (Myerson, 1986). This condition is equivalent to the conditions satisfied by *proper scoring* rules (Savage, 1971; Gneiting and Raftery, 2007).

Implementable rules can be implemented using different mechanisms. One possible implementation allows the analyst to *choose from a restricted set of decision rules*, where each of these rules needs to be monotonic in I, before seeing the data. This corresponds to the actual practice of pre-analysis plans, where the analyst chooses a decision rule before the data becomes available.

The set of implementable rules can be characterized as a *convex polytope*. If the decision-maker objective is convex, and in particular if it is linear, then the optimal implementable rule is necessarily an *extremal point* of this polytope (Vanderbei et al., 2020). Pre-analysis messages allow the decision-maker to implement a larger set of decision rules than would be available without such messages.

Optimal implementable hypothesis tests We next turn to the specific problem of finding optimal implementable hypothesis tests. Such tests are required to satisfy *size control* conditional on both the state of the world and on analyst private information that is available before observing the data. We show that the optimal implementable test, for the decision-maker, can be implemented by (i) requiring the analyst to choose an arbitrary *full-data* test, which is a function of all components X, where this test controls size, and then (ii) implementing this test, making *worst-case assumptions* about any unreported components X_i. The analyst's problem of finding a full-data test that maximizes expected power for this mechanism can again be cast as a linear programming problem. We provide an interactive app that allows the analyst to solve this problem, based on their prior for both data availability J and for the state of the world θ. The output of this app can then serve as a basis for their pre-analysis plan.

Roadmap The rest of this article is structured as follows. We conclude this introduction with a review of some related literature. In Section 2, we present a motivating example concerning statistical testing and p-hacking. In Section 3, we introduce the general model. In Section 4, we characterize implementable decision rules. In Section 5, we characterize optimal implementable hypothesis tests. Appendix A contains all proofs. Appendix B discusses some numerical examples of optimal hypothesis tests subject to the constraint of implementability.

1.1 Related literature

Our article speaks, first, to the current debates around pre-registration – and other possible reforms – in empirical economics and other social- and life-sciences; cf. Christensen and Miguel (2018); Miguel (2021). In doing so, our article applies some of the insights from mechanism design and information design (Myerson, 1986; Kamenica, 2019; Sinander, 2023) to the settings of statistical decision theory and statistical testing, (Wald, 1950; Savage, 1951; Lehmann and Romano, 2006). More broadly, our article contributes to a literature that spans statistics, econometrics and economic theory, and which models statistical inference in multi-agent settings. We differ from other contributions to this literature, in that we focus on the role of implementability as a constraint on statistical decision rules, which rationalizes pre-analysis plans, and on the derivation of optimal decision rules subject to the constraint of implementability.

Drawing on classic references (Tullock, 1959; Sterling, 1959; Leamer, 1974), Glaeser (2006) considers the role of incentives in empirical research. A number of recent contributions model estimation and testing within multiple-agent settings, including Glazer and Rubinstein (2004); Mathis (2008); Chassang et al. (2012); Tetenov (2016); Di Tillio et al. (2021, 2017); Spiess (2018); Henry and Ottaviani (2019); McCloskey and Michaillat (2020); Libgober (2020); Yoder (2020); Williams (2021); Abrams et al. (2021); Viviano et al. (2021). In this literature, Banerjee et al. (2020); Frankel and Kasy (2022); Andrews and Shapiro (2021); Gao (2022) consider the communication of scientific results to an audience with priors, information, or objectives that might differ from the sender's.

The literature on Bayesian persuasion (Kamenica and Gentzkow, 2011; Kamenica, 2019; Curello and Sinander, 2022), like the present article, considers a sender with information unavailable to a receiver, where sender and receiver have divergent objectives. One important way in which our model differs from that of Bayesian persuasion is that in our model the signal space of the analyst is restricted to the truthful but selective reporting of data. This restriction implies that the concavification argument central to Bayesian persuasion does not apply.

2 A motivating example

Before we introduce our general model, consider the following hypothesis-testing problem, where there is a decision-maker and an analyst: The full data consists of two normally distributed components, $X = (X_1, X_2)$, with $X_i \sim \mathcal{N}(\theta, 1)$, independently across components. The X_i might for instance correspond to experimental estimates of an average treatment effect, for two different experimental sites. The analyst observes the subvector X_J. The component $i \in \{1, 2\}$ is observed (i.e., $i \in J$) with probability η_i, again independently across components. Put differently, η_i is the decision-maker's a-priori probability that the analyst successfully implemented an experiment at site i.

The decision-maker is interested in testing the null hypothesis $H_0 : \theta \leq 0$. The analyst, on the other hand, aims to maximize the probability of rejection. The decision-maker does not know which components are actually available, that is, they do not know J. The analyst knows which components are available. Upon learning the data X_J, the analyst chooses the subset $I \subseteq J$, and reports (X_I, I) to the decision-maker, who then rejects the null with probability $A = \mathbf{a}(X_I, I) \in [0,1]$. How should the decision-maker choose the testing rule $\mathbf{a}$ that maps the reported data to a rejection probability?

Five testing rules We compare five different testing rules, $\mathbf{a}_0$ through $\mathbf{a}_4$. For each of these testing rules, Figure 1 shows the rejection probability as a function of (X_1, X_2), assuming that $\eta = (0.9, 0.5)$. This conditional rejection probability given X averages over the availability of components, J. The left panel of Figure 2 shows the corresponding power curves, i.e., the rejection probability as a function of θ, averaging over the distribution of both X and J.

Our benchmark is the **optimal test using all the data**. This test is not, in general, feasible, since not all components are always available. We have that $Z = \frac{1}{\sqrt{2}}(X_1 + X_2) \sim \mathcal{N}(\sqrt{2} \cdot \theta, 1)$ is a sufficient statistic for θ. Since this statistic satisfies the monotone likelihood ratio property, the Neyman–Pearson Lemma implies that the uniformly most powerful test of level α is given by $\mathbf{a}_0(X) = \mathbf{1}(Z > z)$, where $z = \Phi^{-1}(1-\alpha)$; cf. Theorem 3.4.1 in Lehmann and Romano (2006).

Consider next the **naive test** which ignores potentially selective reporting by the analyst. This test acts as if the reported components I are the full data available to the analyst, and implements the corresponding uniformly most powerful test,

$$\mathbf{a}_1(X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{|I|}} \sum_{i \in I} X_i > z\right).$$

The best response of the analyst to this naive testing rule involves selective reporting ("p-hacking"), where $I^* \in \operatorname{argmax}_{I \subseteq J} \mathbf{a}(X_I, I)$. The problem with this naive test is that it does not control size. Selective reporting by the analyst implies that the probability of rejection under the null is not bounded by α.

We might correct for such selective reporting by making worst-case assumptions about all unreported components. This results in the **conservative test**,

$$\mathbf{a}_2(X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{2}}(X_1 + X_2) > z \text{ and } I = \{1,2\}\right).$$

If there are components which are not reported, then the null is not rejected. This conservative test implies a probability of rejection given X of $\eta_1 \cdot \eta_2 \cdot \mathbf{1}\left(\frac{1}{\sqrt{2}}(X_1 + X_2) > z\right)$. The conservative test controls size, but does not have good power properties.

Figure 1: Comparison of testing rules

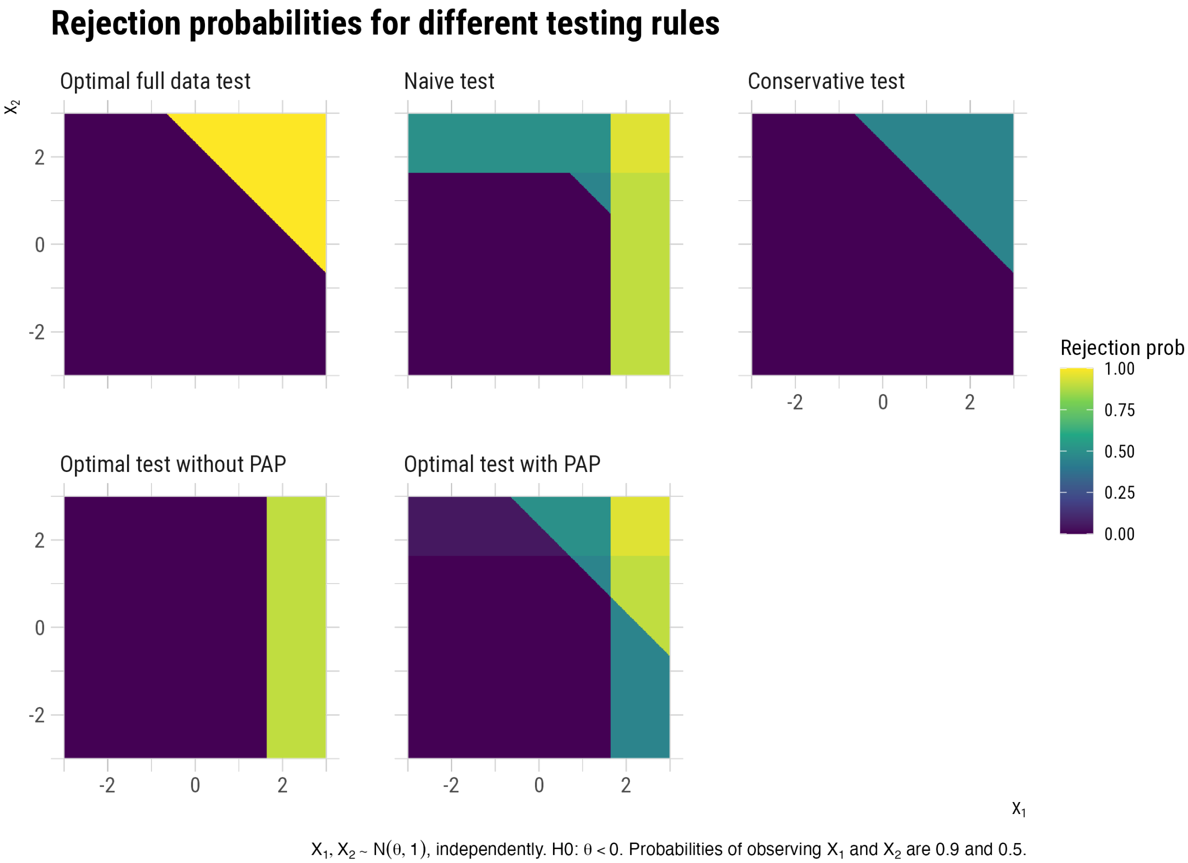
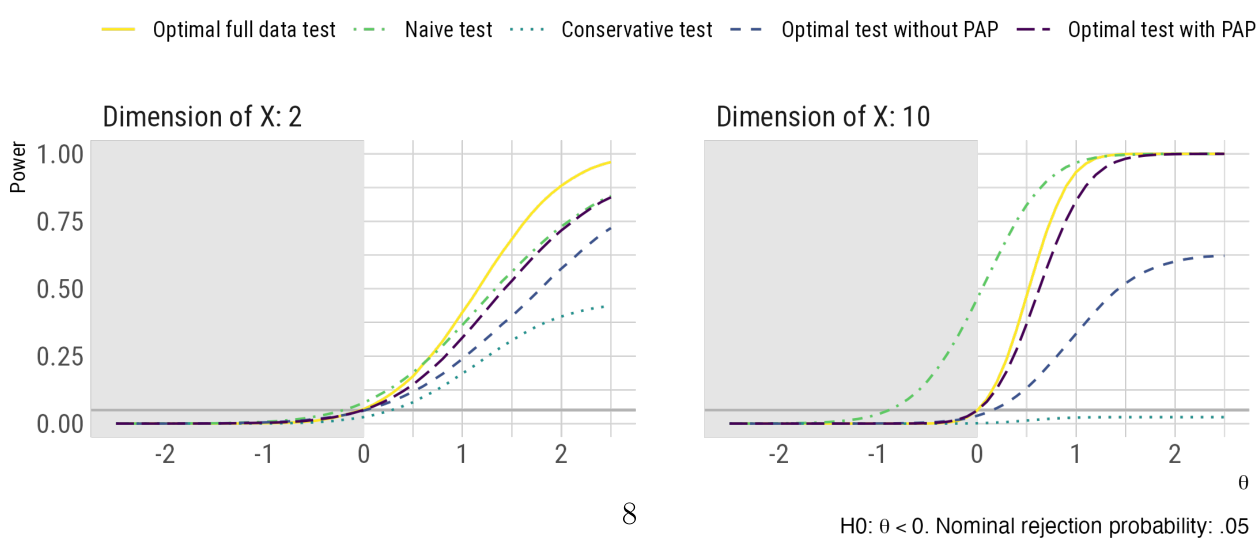


Figure 2: Power curves



As we show more generally in Section 4 and Section 5 below, the **optimal test without a pre-analysis plan** can be implemented by selecting a full-data test of level α. When not all data are reported, the decision-maker needs to assume the worst about the unreported components, and then implements the corresponding full-data test. The decision-maker can choose the full-data test to maximize (ex-ante) expected power, averaging over their prior for θ.

One possible full-data test ignores X_2, which is less likely to be observed in our numerical example, and rejects based on X_1 alone. This results in the test

$$\mathbf{a}_3(X_I, I) = \mathbf{1}\left(X_1 > z \text{ and } 1 \in I\right).$$

This test implies a probability of rejection given X of $\eta_1 \cdot \mathbf{1}\left(X_1 > z\right)$. This test is optimal for some parameter values, while in general the optimal test depends on the prior over θ.[3]

We lastly get to the **optimal test with a PAP**. The optimal test with a PAP is of the same form as the optimal test without a PAP, except that the *analyst* gets to choose the full data test, *prior* to seeing any data. Recall that in our example the analyst knows the components J that are available before possibly reporting a PAP, but we assume that they have no private information regarding θ. (We relax this assumption in our general setup below.) The optimal implementable solution is of the following form. The analyst communicates which components are available as pre-analysis message $M = J$, and the test is given by

$$\mathbf{a}_4(M, X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{|M|}} \cdot \sum_{i \in M} X_i > z \text{ and } M \subseteq I\right).$$

That is, the analyst commits to reporting all components in J, and for that set of components, the most powerful test is implemented.

Comparing size and power The left panel of Figure 2 plots the power curves for the five testing rules, for $n = \dim(X) = 2$, which is the case that we have considered thus far. The right panel shows analogous plots for $n = 10$, with the components of η evenly distributed over a grid from .5 to .9. The latter case illustrates the contrasts

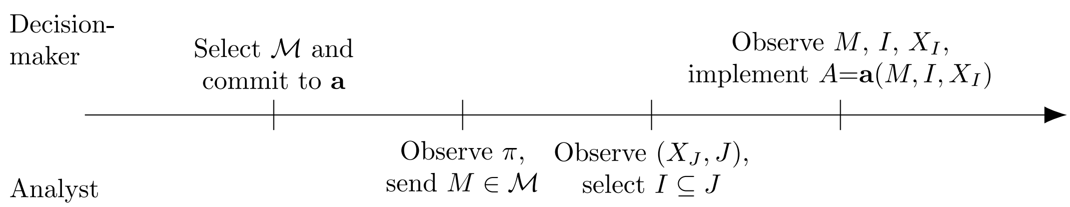
[3]For the given η, this test is for instance optimal when expected power is calculated using the degenerate prior $P(\theta = .3) = 1$.

between testing rules more starkly.

A number of observations are worth emphasizing here. First, the naive test does not control size. For $n = 10$, the probability of rejection for $\theta = 0$ is close to .5, instead of the nominal size of .05. This is due to selective reporting ("p-hacking"). Second, the conservative test can be *very* conservative. Since it only rejects when all components of X are reported, the probability of rejection under the alternative can be arbitrarily small, and remains below the nominal size of .05 for our example with $n = 10$. Third, the optimal test without a PAP does considerably better. It controls size, and is actually strictly conservative under the null. At the same time, it has non-trivial power, which greatly exceeds that of the conservative test. It remains far from optimal, however. The optimal test with a PAP, lastly, controls size exactly under the null. Furthermore, its power under the alternative considerably exceeds that of the optimal test without a PAP.

From our example to the general model Our motivating example is a special case of the general model that we lay out in Section 3. The general model allows for cases where the researcher also has private information about θ, and where the researcher only has partial information about availability J of the data. The general model also covers decision problems other than testing, including estimation and treatment choice.

Figure 3: Timeline



3 Setup

We next describe our general setup, which will be discussed for the rest of this paper. This setup consists of a game between a decision-maker and an analyst. This game is summarized in Assumption 1.[4] The corresponding timeline is shown in Figure 3.

Assumption 1 (Setup)**.** *The game between decision-maker and analyst unfolds as follows:*

1. *The decision-maker selects a message space $\mathcal{M}$ and commits to a decision function $\mathbf{a} : (M, X_I, I) \mapsto A \in \mathcal{A}$.*
2. *The analyst observes the private signal π and sends a message $M \in \mathcal{M}$ to the decision-maker.*
3. *The analyst observes the realization (X_J, J) of available data and selects a subset $I \subseteq J$.*
4. *The decision-maker observes the message M, the subset I, and the data X_I, and implements the decision $A = \mathbf{a}(M, X_I, I)$.*

The analyst and the decision-maker share a common prior P *over the signal π, the parameter θ, the availability J, and the data X. This prior satisfies that the conditional distribution of X only depends on θ, i.e., $X|\theta, J, \pi \overset{d}{=} X|\theta$.*

Discussion This is a game of partial verifiability. The report X_I is always truthful given I, but the non-availability of the components of $\{1, \ldots, k\} \setminus J$ cannot be verified by the decision-maker. *Selective reporting*, where not all available components are

[4]Our notation does not distinguish explicitly between random variables and their realizations. This should not cause any ambiguity. Where the distinction is important, we point this out explicitly.

reported ($I \subsetneq J$), corresponds to p-hacking, or specification searching. Mis-reporting of X_I, which corresponds to scientific fraud, is not allowed in our setting.

The private signal π corresponds to *analyst expertise*. The signal π might be informative about θ, corresponding to knowledge about which hypotheses are likely to be correct, about the likely magnitude of effect sizes, etc. The signal π might also be informative about J, corresponding to knowledge about the viability of different identification approaches, the availability of experimental sites, etc.

There is prior uncertainty of the decision-maker regarding the availability of components J. Without such uncertainty, the mechanism design problem would be trivial, and the decision-maker would simply require the analyst to report everything. Prior uncertainty allows for *"plausible deniability,"* because the decision-maker maker does not know the full set of results from which the reported results were selected.

In Assumption 1, we have left the message space $\mathcal{M}$ for the pre-analysis message M unrestricted. We will later encounter different, equivalent choices for $\mathcal{M}$: The message M might directly communicate the analyst signal π, or their corresponding posterior, in the spirit of the revelation principle in mechanism design. Alternatively, the message M might choose a decision function $\mathbf{a}$ from a restricted set, in the spirit of "aligned delegation" (Frankel, 2014). This latter formulation corresponds more directly to the practice of pre-analysis plans.

Objectives We have not yet described the objectives of either the decision-maker or the analyst. We allow for *conflicting objectives*, which render the mechanism-design problem non-trivial. By contrast, we have imposed *common priors*, so that there are no agency issues driven by divergent beliefs.

We leave the decision-maker's objective unspecified at this point. This allows us to first study implementability as a general constraint on the set of decision-functions available to the decision-maker. This constraint does not depend on the decision-maker objective. We also do not impose that the decision-maker is an expected utility maximizer. This allows us to also study frequentist statistical decision-problems subject to the constraint of implementability, including hypothesis testing subject to size control, and unbiased estimation, in addition to Bayesian decision problems.

By contrast, we do assume that the analyst is an expected utility maximizer. We furthermore impose the following restriction on their utility function for most of our discussion.

Assumption 2 (Monotonic analyst utility)**.** *The analyst is an expected utility maximizer with utility $v(A)$, for a strictly monotonically increasing function v.*

The analyst always prefers a higher outcome $A \in \mathcal{A}$. In the context of testing, the analyst always prefers to reject the null hypothesis. In the context of publication decisions, the analyst always would like their paper to be published. In the context of drug approval, the pharmaceutical company always would like their drug to be approved.

4 Implementability

Conventional statistical decision theory considers decision functions that map the available information into statistical decisions (Wald, 1950; Savage, 1951). In our context, such decision functions $\bar{\mathbf{a}}(\pi, X_J, J)$ map the signal π, the available data X_J, and the component set J into decisions A. We will call such functions $\bar{\mathbf{a}}$ *reduced-form decision functions.*

In our setting, not all such decision functions are available to the decision-maker, because of analyst private information and conflicting objectives. In this section, we will characterize the set of *implementable* reduced form decision functions $\bar{\mathbf{a}}$ which are consistent with analyst utility maximization. This leads to constrained versions of conventional statistical decision problems, including hypothesis testing and point estimation. We will show that implementation, in general, requires the use of pre-analysis messages.

4.1 Which decision functions can be implemented?

The analyst's optimal message M^* and reported set I^* maximize analyst expected utility $\mathrm{E}[v(\mathbf{a}(M, X_I, I))]$, given the decision rule $\mathbf{a}$. Here M^* and I^* are random elements, where M^* is measurable with respect to π, and I^* is measurable with respect to π, X_J, J. Analyst expected utility maximization and strict monotonicity of v imply

$$
\begin{aligned}
I^* &\in \underset{I \subseteq J}{\operatorname{argmax}} \ \mathbf{a}(M^*, X_I, I), \text{ and} \\
M^* &\in \underset{M \in \mathcal{M}}{\operatorname{argmax}} \ \mathrm{E}[v(\mathbf{a}(M, X_{I^*}, I^*))|\pi]. \qquad (1)
\end{aligned}
$$

Consider now reduced-form decision functions $\bar{\mathbf{a}}(\pi, X_J, J)$ that map the information available to the analyst to a decision-maker action. We say that a function $\bar{\mathbf{a}}$ is implementable if it is consistent with analyst utility maximization.

Definition 1 (Implementable reduced-form decision rules)**.** *A reduced form decision function $\bar{\mathbf{a}}(\pi, X_J, J)$ is* implementable *if there exists a decision function* $\mathbf{a}$ *with best responses M^*, I^* such that*

$$\bar{\mathbf{a}}(\pi, X_J, J) = \mathbf{a}(M^*, X_{I^*}, I^*)$$

almost surely.

The following theorem provides a complete characterization of implementable reduced-form decision rules in our setting. The proof of this theorem, and all subsequent proofs, can be found in Appendix A.[5]

Theorem 1 (Implementability)**.** *Under Assumptions 1 and 2, a reduced-form decision function $\bar{\mathbf{a}}(\pi, X_J, J)$ is implementable if and only if there is some $\tilde{\mathbf{a}}$ such that $\bar{\mathbf{a}}(\pi, X_J, J) = \tilde{\mathbf{a}}(\pi, X_J, J)$ almost surely, and both of the following two conditions hold:*

1. ***Truthful message:*** *For all π, π',*

$$\mathrm{E}[v(\tilde{\mathbf{a}}(\pi', X_J, J))|\pi] \leq \mathrm{E}[v(\tilde{\mathbf{a}}(\pi, X_J, J))|\pi]. \tag{2}$$

2. ***Monotonicity:*** *For all π, X, J and $I \subseteq J$,*

$$\tilde{\mathbf{a}}(\pi, X_I, I) \leq \tilde{\mathbf{a}}(\pi, X_J, J). \tag{3}$$

Theorem 1 characterizes which reduced-form decision functions $\bar{\mathbf{a}}(\pi, X_J, J)$ can be implemented, but it does not tell us *how* to implement them. The following Proposition 1 shows two different, canonical ways of implementing any such function. The first implementation uses truthful revelation of analyst signals. The second implementation uses delegation, where the analyst is effectively allowed to choose the

[5]It is worth noting that the revelation principle (Myerson, 1986) does not directly apply to our setting, since misreporting of analyst "types" is constrained by the verifiability of their reports (X_I, I), and by $I \subseteq J$. See Kephart and Conitzer (2016) for a discussion of the revelation principle under partial verifiability and, more generally, for settings where misreporting is potentially costly.

decision function from a pre-specified, restricted set $\mathcal{B}$. This second implementation corresponds closely to the actual practice of pre-analysis plans. Here, the analyst pre-specifies a mapping b from the reported data (X_J, J) to the decision $A = b(X_J, J)$. Proposition 1 shows that restricting attention to implementation by such pre-analysis plans is without loss of generality.

Proposition 1 (Implementation)**.** *Under Assumptions 1 and 2, a reduced-form decision rule* $\bar{\mathbf{a}}$ *can be implemented if and only if either of the following two conditions holds:*

1. ***Implementation by truthful revelation:*** $\bar{\mathbf{a}}$ *can be implemented with a decision rule* $\mathbf{a}$ *for which*

$$\mathbf{a}(\pi, X_J, J) = \bar{\mathbf{a}}(\pi, X_J, J),$$

where the message space is the set of all possible signals π*.*

2. ***Implementation by delegation (pre-analysis plan):*** $\bar{\mathbf{a}}$ *can be implemented with a decision rule* $\mathbf{a}$ *for which*

$$\mathbf{a}(b, X_J, J) = b(X_J, J),$$

where b *is restricted to lie in some set* $\mathcal{B}$ *that acts as the message space.*

4.2 Alternative characterizations of implementability

Having characterized implementable decision functions in general, we next discuss implementability for the special case of linear analyst utility v and convex action space $\mathcal{A}$. We also discuss the connection of truthful revelation to proper scoring, as well as possible simplicity restrictions on pre-analysis plans.

The set of implementable rules as a convex polytope In addition to Assumptions 1 and 2, assume for a moment that the action space $\mathcal{A} \subseteq \mathbb{R}$ is convex, and that analyst utility is linear – without additional loss of generality, $v(A) = A$. The leading examples involve binary decisions, where we interpret A as the *probability* of a positive decision. Binary decisions occur for statistical testing, as discussed in Section 5 below, as well as for publication decisions, drug approval, etc. Linearity is without loss of generality for the case of binary decisions; in this case, it follows from expected utility maximization. Suppose finally that π has finite support.

Under these additional assumptions, we get that the set of implementable reduced form decision functions $\bar{\mathbf{a}}$ is given by a convex polytope, characterized by the following constraints.

$$\bar{\mathbf{a}}(\pi, X_J, J) \in \mathcal{A}, \tag{Support}$$

$$\bar{\mathbf{a}}(\pi, X_I, I) - \bar{\mathbf{a}}(\pi, X_J, J) \leq 0 \quad \forall\, \pi, X_J, J, I \subseteq J, \tag{Monotonicity}$$

$$\sum_{X_J, J} \left(\bar{\mathbf{a}}(\pi', X_J, J) - \bar{\mathbf{a}}(\pi, X_J, J)\right) \mathrm{P}_\pi(X_J, J) \leq 0 \quad \forall\, \pi', \pi. \tag{Truthful message}$$

In the last inequality, P_π is a shorthand for the analyst's posterior distribution conditional on π.

If, furthermore, the decision-maker objective is linear in $\bar{\mathbf{a}}$, as is the case for a Bayesian decision-maker and binary actions, or if it is linear with an additional linear constraint, as is the case for expected power maximization subject to size control, then the problem of finding the optimal implementable reduced form decision function becomes a linear programming problem. Efficient algorithms exist for numerically solving such problems, cf. Vanderbei et al. (2020). We will return to this point in Section 5 below. We leverage such linear programming algorithms in our interactive app for finding optimal PAPs.

Truthful revelation of beliefs and proper scoring Condition (2) in Theorem 1 ensures that the analyst reveals their relevant prior information truthfully. Condition (2) is equivalent to the definition of a proper scoring rule, as introduced by Savage (1971). The theory of proper scoring rules has regained importance in the more recent statistics and machine learning literature, cf. Gneiting and Raftery (2007).

Given a reduced form decision rule $\bar{\mathbf{a}}$, define

$$S(\pi', \pi) = \mathrm{E}_\pi[v(\bar{\mathbf{a}}(\pi', X_J, J))]. \tag{4}$$

The expectation E_π is taken over the conditional prior distribution P_π of X_J, J given π. Denote the Euclidean inner product for functions of X_J, J (understood here as values, rather than as random variables) by $\langle f(\cdot), g(\cdot)\rangle = \sum_{X_J,J} f(X_J, J) \cdot g(X_J, J)$. Here we assume for simplicity that X has finite support, though the argument generalizes. We obtain the following characterization, which was first stated by Savage (1971) and is restated as Theorem 2 in Gneiting and Raftery (2007).

Proposition 2 (Proper scoring rule)**.** *Condition* (2)*, the truthful message condition, holds for all* π, π' *if and only if there exists a convex function* G *with sub-gradient* G'*, where* $G(\mathrm{P}_\pi) = S(\pi, \pi)$ *on the support of* π*, such that* $S(\pi', \pi) = G(\mathrm{P}_{\pi'}) + \langle G'(\mathrm{P}_{\pi'}, \cdot), \mathrm{P}_\pi - \mathrm{P}_{\pi'} \rangle$.

Simple pre-analysis plans Item 2 of Proposition 1 shows that reduced form decision rules can be implemented by delegation: The decision-maker offers a set $\mathcal{B} = \{b : (X_I, I) \mapsto \mathcal{A}\}$ of permissible pre-analysis plans (decision functions). The analyst then commits to one of the decision functions $b \in \mathcal{B}$ before access to the data.

In practice, some pre-analysis plans may be unrealistically complicated, and we may wish to restrict attention to a smaller set $\mathcal{B}_0$ of simpler mappings. The decision-maker's choice would then be restricted to $\mathcal{B} \subseteq \mathcal{B}_0$ as a subset of feasible mappings.

4.3 Are pre-analysis messages needed?

Aligned objectives Why does implementability in our setting appear to require a pre-analysis message, if that is not the case in conventional statistical decision theory? Assume for a moment that analyst and decision-maker share the same objective function. In this case, is there any need for a pre-analysis message? The answer is no.

To see this, consider the following variant of our setup. Suppose everything is as in Assumption 1 (Figure 3), except that the analyst gets to choose the message M *after* they observe the data X_J, J. Put differently, the analyst cannot provide a verifiable time-stamp for their message M to the decision-maker. The following observation states that in this modified setting, where there is no *pre*-analysis message, the decision-maker can still implement the first-best reduced-form decision rule, provided that preferences are aligned.

Proposition 3 (First-best decisions for aligned preferences)**.** *Under the modified Assumption 1 where the message* M *can depend on the realization* (X_J, J)*, assume that analyst and decision-maker are expected utility maximizers who share the same utility function* $u(A, \theta)$*. Then the decision-maker's first-best reduced-form decision rule* $\bar{\mathbf{a}}(\pi, X_J, J)$ *is implementable.*

As Proposition 3 shows, *pre*-analysis messages only become potentially useful in the presence of both private information *and* misaligned preferences.

Implementability without pre-analysis message We next characterize the set of decision functions $\bar{\mathbf{a}}$ that are implementable without a pre-analysis message, when objectives can be misaligned. In this case, the implementable functions are exactly the functions $\bar{\mathbf{a}}(\pi, X_J, J)$ that satisfy monotonicity with respect to set inclusion for the index set J given X, and that do not depend on π. Analyst expertise can thus not be used to improve decisions *at all*, in the absence of a pre-analysis message. The proof of the following proposition parallels the proof of Theorem 1.

Proposition 4 (Implementability without pre-analysis message)**.** *Under Assumptions 1 and 2, with the additional constraint that there is no pre-analysis message, a reduced-form decision function $\bar{\mathbf{a}}$ is implementable if and only if there is a function $\tilde{\mathbf{a}}$ with almost surely $\bar{\mathbf{a}}(\pi, X_J, J) = \tilde{\mathbf{a}}(X_J, J)$ and*

$$\tilde{\mathbf{a}}(X_I, I) \leq \tilde{\mathbf{a}}(X_J, J) \tag{5}$$

for almost all X, J and all $I \subseteq J$.

5 Frequentist hypothesis testing

We next specialize our general framework to the setting of frequentist hypothesis testing. In this setting, the decision-maker decides whether to reject a null hypothesis. We assume that the decision-maker wants to maximize expected power subject to size control. The analyst, however, always prefers a rejection of the null hypothesis.

Building on our previous results, we characterize the set of implementable testing rules that satisfy size control, in Section 5.2. We provide a simple mechanism that allows the decision-maker to implement the optimal testing rule. This mechanism requires a pre-analysis plan, where the analyst may choose any full-data test that satisfies size control, and the decision-maker then makes worst-case assumptions about any unreported data. This mechanism solves the decision-maker's problem.

In Section 5.3 we then consider the analyst's problem of finding an optimal response to this mechanism, and show that they have to solve a linear programming problem to find the optimal pre-analysis plan. We provide software to solve this problem of the analyst. We also characterize the set of possible solutions to the analyst's problem, by describing the set of extremal points of their feasible set.

Throughout, we focus on the problem of testing a single (joint) hypothesis, and leave an extension to deciding which of multiple hypotheses to reject for future work.

5.1 Decision-maker and analyst objectives

Assume that the decision $A \in [0,1]$ represents the probability, given (M, X_J, J), of rejecting the null hypothesis $\theta \in \Theta_0$. Suppose that the analyst is an expected utility maximizer, who ex-post only cares about the binary testing decision. Ex-ante, the analyst thus wants to maximize expected power. It follows that their utility is linear in A. We can then make the following normalizing assumption, without loss of generality.

Assumption 2' (Power analyst utility)**.** *Analyst utility is*

$$v(A) = A.$$

The decision-maker also wants to maximize expected power, but subject to the constraint of size control under the null hypothesis.

Definition 2 (Size control)**.** *We say that a reduced-form decision rule* $\bar{\mathbf{a}}$ *controls size at level* $\alpha \in (0,1)$ *if*

$$\sup_{\pi, \theta \in \Theta_0, J \subseteq \{1,\ldots,n\}} \mathrm{E}[\bar{\mathbf{a}}(\pi, X_J, J) | \theta, \pi, J] \leq \alpha. \tag{6}$$

Recall that we imposed, in Assumption 1, that the conditional distribution of X only depends on θ, that is, $X|\theta, J, \pi \overset{d}{=} X|\theta$. Under this assumption, the conditional expectation $\mathrm{E}[\bar{\mathbf{a}}(\pi, X_J, J)|\theta, \pi, J]$ is well-defined even outside the joint support of π, θ, J, as long as θ is within its marginal support.

5.2 Decision-maker solution: Pre-specified full-data tests

The implementability results of Section 4 allow us to characterize optimal pre-analysis plans for hypothesis testing as follows.

Theorem 2 (Optimal pre-analysis plans with size control)**.** *Define $\mathcal{T}$ to be the class of measurable full-data tests $t : \mathcal{X} \to [0, 1]$ satisfying size control, $\sup_{\theta \in \Theta_0} \mathrm{E}[t(X)|\theta] \leq \alpha$. Under Assumption 1 and Assumption 2', the power-maximizing decision rule subject to the constraints of implementability (Definition 1) and size control (Definition 2) can be implemented by requiring the analyst to communicate, as a pre-analysis message, a full-data test $t \in \mathcal{T}$, and then rejecting the null with conditional probability*

$$b(X_I, I) = \inf_{X';\, X'_I = X_I} t(X').$$

This result builds on the general characterizations of Theorem 1 and Proposition 1. To get further intuition for Theorem 2 note, first, that it is sufficient to verify size control for the *full-data* test t. The reason is that implementable reduced-form decision rules must fulfill the monotonicity constraint (3). Subject to monotonicity in I, size control of $\bar{\mathbf{a}}$ in the sense of Definition 2 is equivalent to size control for the full-data test $\bar{\mathbf{a}}(\pi, X, \{1, \ldots, k\})$.

Note, second, that for *optimal* reduced-form testing rules the monotonicity constraint is in general binding, since both decision-maker and analyst aim to maximize expected power, subject to the constraints. For optimal rules it is therefore without loss of generality to assume $\bar{\mathbf{a}}(\pi, X_J, J) = \inf_{X';\, X'_J = X_J} t(X')$, which can be implemented by b as in the statement of the theorem.

5.3 Analyst solution: Linear programming

Theorem 2 solves the optimal testing problem from the decision-maker's perspective: Let the analyst pre-specify a valid full-data test, and then make worst-case assumptions about unreported data. We next turn to the analyst's problem: What full-data test should they specify? This problem can be cast as a linear programming problem. The optimal value for any linear programming problem can be achieved on the set of extremal points of the feasible set.[6] This insight, which is of central importance to mechanism design (Sinander, 2023), allows us to characterize the set of potential solutions to the optimal testing problem subject to implementability.

Linear objective and linear feasible set For ease of exposition, we focus on point null hypotheses $\Theta_0 = \{\theta_0\}$ in the following. Our results easily extend to com-

[6]The same holds more generally, for the maximum of a convex function on a convex set.

pound hypotheses. Denote $K = \{1, \dots, k\}$ the index set of all potentially available components. Let $\mathcal{B}$ be the set of measurable functions $b(X_J, J)$ defined by the following constraints.

$$\int b(X, K) d\,\mathrm{P}_{\theta_0}(X) \leq \alpha, \qquad \text{(Size control)}$$

$$b(X_J, J) \in [0, 1] \qquad \forall\, J, X, \qquad \text{(Support)}$$

$$b(X_J, J) \leq b(X, K) \qquad \forall\, J, X. \qquad \text{(Monotonicity)}$$

This is the set of testing rules from which the analyst is effectively allowed to choose, after observing their private signal π. This characterization applies to both discrete and continuously distributed X. The set $\mathcal{B}$ is a convex polytope.

The (interim) analyst objective function is given by expected power, conditional on their private signal π,

$$\mathrm{E}_\pi[b(X_J, J)] = \int b(X_J, J) d\,\mathrm{P}_\pi(X, J). \qquad \text{(Interim expected power)}$$

We provide code, in the form of an interactive app, which allows the analyst to easily solve the problem of maximizing expected power, subject to $b \in \mathcal{B}$.[7]

Potentially optimal tests: Extremal points of $\mathcal{B}$ Suppose we maintain Assumption 1 and Assumption 2', but impose no further assumptions on the (interim) prior P_π of the analyst. What can we say about the set of potential solutions b to the analyst's problem, in this case? The following proposition provides a characterization, based on the set of extremal points of the set $\mathcal{B}$, intersected with the set of rules b for which monotonicity is binding.

Proposition 5. *Suppose that Assumption 1 and Assumption 2' hold, and consider the mechanism specified in Theorem 2. Then there exists a full-data test t which is a best response of the analyst such that $b(X_J, J) = \inf_{X' \colon X'_J = X_J} t(X')$ is extremal in $\mathcal{B}$. Suppose that t takes on a finite number of values. Then a function b of this form is extremal in $\mathcal{B}$ if and only if the following conditions hold:*

1. *$t(X) \in \{0, q, 1\}$ for all X, for some $0 < q < 1$.*
2. *If there exists X such that $t(X) = q$, then $\mathrm{P}_{\theta_0}(t(X) = q) > 0$.*

[7] This app is available at `https://maxkasy.shinyapps.io/The_PAP_App/`.

3. *For any $X \neq X'$ such that $t(X) = t(X') = q$, there exists a value J such that $X_J = X'_J$ and $b(X_J, J) = b(X'_J, J) = q$.*

In other words, we can restrict our attention to testing rules that partition values of the data X into at most three regions: one where the test always rejects; one where the test never rejects; and one where it rejects with a single, intermediate probability. Furthermore, if there is more than one value for which the test takes this intermediate rejection probability, then the monotonicity constraint in the construction of the tests b is binding for at least some subset J.

The result in Proposition 5 characterizes the set of extremal points of $\mathcal{B}$ for which monotonicity is binding. The optimal analyst response is necessarily in this set. Can all of these points be rationalized as optimal for some analyst interim prior? The following proposition provides a partial answer.

Proposition 6. *Suppose that $\mathrm{P}_{\theta_0}(b(X, K) \notin \{0, 1\}) = 0$ for $b \in \mathcal{B}$. Then there exists a prior $\mathrm{P}_\pi(X_J, J)$ such that b maximizes the objective $\int b(X_J, J) d\,\mathrm{P}_\pi(X_J, J)$ in $\mathcal{B}$.*

This result shows that all testing rules that control size without an intermediate probability of rejection can be rationalized.

References

Abrams, Eliot, Jonathan Libgober, and John A List (2021). Research registries and the credibility crisis: An empirical and theoretical investigation.

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

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey (2023). Inference on Winners. *The Quarterly Journal of Economics*.

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

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

Chassang, Sylvain, Gerard Padró I Miquel, and Erik Snowberg (2012). Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments. *The American Economic Review*, 102(4):1279–1309.

Christensen, Garret and Edward Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Coffman, Lucas C. and Muriel Niederle (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, 29(3):81–98.

Curello, Gregorio and Ludvig Sinander (2022). The comparative statics of persuasion. *arXiv preprint arXiv:2204.07474*.

Di Tillio, Alfredo, Marco Ottaviani, and Peter Norman Sørensen (2017). Persuasion bias in science: Can economics help? *Economic Journal*, 127(605):266–304.

Di Tillio, Alfredo, Marco Ottaviani, and Peter Norman Sørensen (2021). Strategic sample selection. *Econometrica*, 89(2):911–953.

Duflo, Esther, Abhijit V Banerjee, Amy Finkelstein, Lawrence F Katz, Benjamin Olken, and Anja Sautmann (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics. *NBER Working Paper*, (w26993).

Food and Drug Administration (1998). Guidance for industry: Statistical principles for clinical trials. *US Department of Health and Human Services.*

Frankel, Alexander (2014). Aligned delegation. *American Economic Review*, 104(1):66–83.

Frankel, Alexander and Maximilian Kasy (2022). Which findings should be published? *American Economic Journal: Microeconomics*, 14(1):1–38.

Gao, Ying (2022). Inference from selectively disclosed data. *arXiv preprint arXiv:2204.07191.*

Glaeser, Edward L (2006). Researcher Incentives and Empirical Methods. Technical Report t0329, National Bureau of Economic Research.

Glazer, Jacob and Ariel Rubinstein (2004). On optimal rules of persuasion. *Econometrica*, 72(6):1715–1736.

Gneiting, Tilmann and Adrian E Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Henry, Emeric and Marco Ottaviani (2019). Research and the Approval Process: The Organization of Persuasion. *American Economic Review*, 109(3):911–955.

Kamenica, Emir (2019). Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272.

Kamenica, Emir and Matthew Gentzkow (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kephart, Andrew and Vincent Conitzer (2016). The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 85–102.

Leamer, Edward E (1974). False Models and Post-Data Model Construction. *Journal of the American Statistical Association*, 69(345):122–131.

Lehmann, Erich L and Joseph P Romano (2006). *Testing statistical hypotheses.* Springer Science & Business Media.

Libgober, Jonathan (2020). False Positives and Transparency.

Mathis, Jérôme (2008). Full revelation of information in sender–receiver games of persuasion. *Journal of Economic Theory*, 143(1):571–584.

McCloskey, Adam and Pascal Michaillat (2020). Incentive-Compatible Critical Values. Technical Report 2005.04141.

Miguel, Edward (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives*, 35(3):193–214.

Myerson, Roger B (1986). Multistage games with communication. *Econometrica: Journal of the Econometric Society*, pages 323–358.

Olken, Benjamin A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Savage, L J (1951). The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67.

Savage, Leonard J (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Sinander, Ludvig (2023). Topics in mechanism design. *Lecture notes.*

Spiess, Jann (2018). Optimal estimation when researcher and social preferences are misaligned.

Sterling, Theodore 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.

Tetenov, Aleksey (2016). An economic theory of statistical testing. Technical Report CWP50/16, cemmap working paper.

Tullock, Gordon (1959). Publication Decisions and Tests of Significance—A Comment. *Journal of the American Statistical Association*, 54(287):593–593.

Vanderbei, Robert J et al. (2020). *Linear programming*. Springer.

Viviano, Davide, Kaspar Wuthrich, and Paul Niehaus (2021). (When) should you adjust inferences for multiple hypothesis testing?

Wald, Abraham (1950). *Statistical decision functions*. Wiley New York.

Williams, Cole (2021). Preregistration and Incentives.

Yoder, Nathan (2020). Designing Incentives for Heterogeneous Researchers.

A Proofs

Implementability

Proof of Theorem 1.
We first show that existence of such an $\tilde{\mathbf{a}}$, which satisfies conditions (2) and (3), implies implementability. We then show that implementability implies existence of such an $\tilde{\mathbf{a}}$.

Assume first that such an $\tilde{\mathbf{a}}$ exists. Then, letting the message space be the space of signals π, and choosing $\mathbf{a}(\pi, X_I, I) = \tilde{\mathbf{a}}(\pi, X_I, I)$, yields incentive compatibility of $I^* = J, M^* = \pi$: For any alternative π, X_J, J-measurable reporting policy $\tilde{I} \subseteq J$ and message $\tilde{M} = \pi'$, we have that

$$\begin{aligned} v(\mathbf{a}(M^*, \tilde{I}, X_{\tilde{I}})) &\leq v(\mathbf{a}(M^*, I^*, X_{I^*})) \\ \mathrm{E}[v(\mathbf{a}(\tilde{M}, \tilde{I}, X_{\tilde{I}}))|\pi] &\leq \mathrm{E}[v(\mathbf{a}(\pi', J, X_J))|\pi] \\ &\leq \mathrm{E}[v(\mathbf{a}(\pi, J, X_J))|\pi] = \mathrm{E}[v(\mathbf{a}(M^*, I^*, X_{I^*}))|\pi] \end{aligned}$$

The first inequality holds by monotonicity of $\tilde{\mathbf{a}}$. The first inequality in the second line also holds by monotonicity of $\tilde{\mathbf{a}}$. The last inequality holds because of the truthful message condition. For this choice of I^*, M^*, we have $\bar{\mathbf{a}}(\pi, X_J, J) = \tilde{\mathbf{a}}(\pi, X_J, J)$ almost surely, as desired.

Assume now reversely that the reduced-form decision function $\bar{\mathbf{a}}$ is implementable by a decision rule $\mathbf{a}$, with π, X_J, J-measurable analyst choices I^* and π-measurable analyst message $M^* = M^*(\pi)$. Define

$$\tilde{\mathbf{a}}(\pi, X_J, J) = \max_{I \subseteq J} \mathbf{a}(M^*(\pi), X_I, I).$$

Note that $\tilde{\mathbf{a}}$ is also well-defined for values of π, X_J, J outside the joint support of these variables. By definition of the reduced form policy, we immediately get

$$\bar{\mathbf{a}}(\pi, X_J, J) = \tilde{\mathbf{a}}(\pi, X_J, J)$$

almost surely (i.e., on the joint support of π, X_J, J).

To see that $\tilde{\mathbf{a}}(\pi, X_J, J)$ satisfies monotonicity note that the maximum over I can only increase, when it is taken over a larger set of possible values for the set of components I. To see that $\tilde{\mathbf{a}}(\pi, X_J, J)$ also satisfies the truthful message condition, note that

$$\begin{aligned}
\mathrm{E}[v(\tilde{\mathbf{a}}(\pi, X_J, J))|\pi] &= \mathrm{E}[\max_{I\subseteq J} v(\mathbf{a}(M^*(\pi), X_I, I))|\pi] \\
&= \max_{M\in\mathcal{M}} \mathrm{E}[\max_{I\subseteq J} v(\mathbf{a}(M, X_I, I))|\pi] \\
&\geq \mathrm{E}[\max_{I\subseteq J} v(\mathbf{a}(M^*(\pi'), X_I, I))|\pi] \\
&= \mathrm{E}[v(\tilde{\mathbf{a}}(\pi', X_J, J))|\pi].
\end{aligned}$$

The first equality holds given the definition of $\tilde{\mathbf{a}}$. The second equality holds given the definition incentive compatibility for $M^*(\pi)$. The following inequality holds since the maximum over M is necessarily weakly larger than the value for any given message $M^*(\pi')$. The last equality, finally, again holds given the definition of $\tilde{\mathbf{a}}$. The claim follows. $\square$

Proof of Proposition 1.
The first part follows from the arguments in the proof of Theorem 1, where we set $\mathbf{a}(\pi, X_I, I) = \tilde{\mathbf{a}}(\pi, X_I, I)$. Note, in particular, that if a rule is implementable using a π-measurable message $M^*(\pi)$, then it is also implementable with the signal π itself as the message, via the decision rule $\mathbf{a}(\pi, X_I, I) = \mathbf{a}'(M^*(\pi), X_I, I)$.

For the second alternative, implementation using delegation, assume first that $\bar{\mathbf{a}}$ is implementable by some decision rule $\mathbf{a}$ with message space $\mathcal{M}$. Then it is implementable by offering the analyst a choice from $\mathcal{B} = \{(X_I, I) \mapsto \mathbf{a}(M, X_I, I); M \in \mathcal{M}\}$. Assume reversely that $\bar{\mathbf{a}}$ is implementable by the proposed delegation mechanism. Then it is implementable by the decision rule $\mathbf{a}(b, X_I, I) = b(I, X_I)$ with message space $\mathcal{M} = \mathcal{B}$. $\square$

Proof of Proposition 2.
The following is based on the proof of Theorem 1 (a generalization of Savage's theorem) in Gneiting and Raftery (2007). A scoring rule is called proper if it satisfies Condition (2), the truthful message condition.

We first show that the characterization in the proposition is sufficient for the scoring rule S to be proper. Convexity of G and the definition of S based on G immediately imply that S is proper, i.e., that truthful revelation is incentive compatible, since convexity implies

$$S(\pi,\pi) = G(\mathrm{P}_\pi) \geq G(\mathrm{P}'_\pi) + \langle G'(\mathrm{P}_{\pi'},\cdot), \mathrm{P}_\pi - \mathrm{P}_{\pi'}\rangle = S(\pi',\pi),$$

for any subgradient G'.

Reversely, suppose that $S(\pi',\pi)$ is a proper scoring rule. Linearity in P_π holds by definition, since $S(\pi',\pi)$ is defined, in (4), as an expectation over P_π. $S(\pi',\pi)$ is thus, in particular, a convex function of P_π. $G(\mathrm{P}_\pi) = S(\pi,\pi) = \sup_{\pi'} S(\pi',\pi)$ is an upper envelope of convex functions, and therefore convex itself. Furthermore, $S(\pi',\cdot)$ is a subgradient of G at π' by definition of proper scoring rules. The claim follows. $\square$

Proof of Proposition 3.
Denote by

$$\tilde{\mathbf{a}}(\pi, X_J, J) = \underset{A\in\mathcal{A}}{\operatorname{argmax}}\ \mathrm{E}[u(a,\theta)|\pi, X_J, J]$$

the first-best reduced-form decision rule of the decision-maker. Let $\mathcal{M}$ be the set of all signals π, and choose $\mathbf{a}$ such that $\mathbf{a}(\pi, I, X_I) = \tilde{\mathbf{a}}(\pi, X_I, I)$. In this case, $M^* = \pi$ and $I^* = J$ are best responses that implement $\tilde{\mathbf{a}}$. $\square$

Proof of Proposition 4.
Suppose first that the monotonicity condition (5) holds. Then $\mathbf{a}(X_I, I) = \tilde{\mathbf{a}}(X_I, I)$ yields incentive compatibility of $I^* = J$, since for any alternative π, X_J, J-measurable reporting policy $\tilde{I} \subseteq J$ we have that

$$v(\mathbf{a}(\tilde{I}, X_{\tilde{I}})) \leq v(\mathbf{a}(I^*, X_{I^*})).$$

by monotonicity of $\mathbf{a}$. For this choice of I^*, $\bar{\mathbf{a}}(\pi, X_J, J) = \tilde{\mathbf{a}}(X_J, J)$ almost surely, as desired.

Conversely, consider an arbitrary decision function $\bar{\mathbf{a}}$ that is implementable by a decision rule $\mathbf{a}$ and π, X_J, J-measurable analyst choice I^*. Since I^* is an analyst best-response to this decision function $\mathbf{a}$, it follows that the corresponding reduced form decision function satisfies

$$\bar{\mathbf{a}}(\pi, X_J, J) = \mathbf{a}(X_{I^*}, I^*) = \max_{I \subseteq J} \mathbf{a}(X_I, I)$$

almost surely. The right-hand side does not depend on π, and the maximum (weakly) increases whenever the maximum is taken over a larger set of possible values for I. The monotonicity condition (5) follows for $\tilde{\mathbf{a}}(X_J, J) = \max_{I \subseteq J} \mathbf{a}(X_I, I)$, which is defined for arbitrary J. $\square$

Hypothesis testing

Proof of Theorem 2:
The mechanism described in Theorem 2 corresponds to the second characterization of implementability in Proposition 1. Define $\tilde{\mathcal{B}}$ as the set of functions b of the form

$$b(X_J, J) = \inf_{X';\, X'_J = X_J} t(X'),$$

for some full-data tests $t : \mathcal{X} \to [0, 1]$ satisfying size control, $\sup_{\theta \in \Theta_0} \mathrm{E}[t(X)|\theta] \leq \alpha$. This $\tilde{\mathcal{B}}$ is the set of decision functions from which the analyst can effectively choose at the pre-analysis stage.

For any such b, monotonicity of $b(X_J, J)$ is immediate. Monotonicity of b and size control of t implies, together with $X|\theta, \pi, J \stackrel{d}{=} X|\theta$ from Assumption 1, that

$$\mathrm{E}[b(X_J, J)|\theta, \pi, J] \leq \mathrm{E}[t(X)|\theta, \pi, J] = \mathrm{E}[t(X)|\theta] \leq \alpha,$$

for all $\theta \in \Theta_0$, so that b satisfies size control.

It remains to show that the b chosen by the analyst has maximal expected power among all decision functions satisfying size control and monotonicity. Since the analyst aims to maximize expected power, it suffices to show that for any $\tilde{b}$ which satisfies size control and monotonicity, the set $\tilde{\mathcal{B}}$ contains a decision function b with power at least as high as that for $\tilde{b}$.

To see that this is the case, take any $\tilde{b}$ satisfying size control and monotonicity. Define $t(X) = \tilde{b}(X, \{1, \ldots, k\})$, and define $b(X_J, J) = \inf_{X';\, X'_J = X_J} t(X')$. Then

$b(X_J, J) \geq \tilde{b}(X_J, J)$ for all X_J, J, and $b \in \tilde{\mathcal{B}}$. In particular, expected power for b is at least as high as for $\tilde{b}$. The claim follows. $\square$

To prove Proposition 5, note first that an element of $\mathcal{B}$ is extremal if and only if there exists no function $\Delta = \Delta(X_J, J)$, where $\Delta \not\equiv 0$, such that both $b+\Delta$ and $b-\Delta$ lies in $\mathcal{B}$.

Lemma 1. *Suppose that $b \in \mathcal{B}$. Then $b + \Delta \in \mathcal{B}$ and $b - \Delta \in \mathcal{B}$ if and only if the following conditions hold:*

$$\int \Delta(X, K) d\,\mathrm{P}_{\theta_0}(X) = 0 \tag{7}$$

$$|\Delta(X_J, J)| \leq \min(b(X_J, J), 1 - b(X_J, J)) \qquad \forall\ J, X \tag{8}$$

$$|\Delta(X_J, J) - \Delta(X, K)| \leq b(X, K) - b(X_J, J) \qquad \forall\ J, X. \tag{9}$$

Proof of Lemma 1:
Immediate. Each of the three conditions corresponds to one of the conditions defining $\mathcal{B}$ (size control, support, and monotonicity). $\square$

Proof of Proposition 5:
The first part of the proposition is immediate from our preceding discussion; we prove the characterization of extremal points. We first show that the stated conditions are sufficient for b to be extremal.
Suppose Δ satisfies the conditions of Lemma 1, and b satisfies the conditions of this proposition. We need to show that $\Delta \equiv 0$.

1. By condition (8), $\Delta(X, K) = 0$ for all X such that $b(X, K) \in \{0, 1\}$.

2. If there exists no X such that $b(X, K) = q$, it follows that $\Delta(X, K) = 0$ for all X.

3. If there exists only one X such that $b(X, K) = q$, we denote $\Delta(X, K) = \delta$.

   If there exist two points $X \neq X'$ such that $b(X, K) = b(X', K) = q$, then by assumption there is also some J such that $b(X, K) = b(X', K) = b(X_J, J) = b(X'_J, J) = q$ and $X_J = X'_J$. Condition (9) then implies $\Delta(X, K) = \Delta(X_J, J) =$

$\Delta(X', K)$. $\Delta(X, K)$ is therefore constant for all X such that $b(X, K) = q$. Write $\Delta(X, K) = \delta$ for such values of X.

It follows that $\int \Delta(X, K) d\,\mathrm{P}_{\theta_0}(X) = \delta \cdot \mathrm{P}_{\theta_0}(b(X, K) = q)$.

4. Condition (7), in combination with $\mathrm{P}_{\theta_0}(b(X, K) = q) > 0$ if there exists any X such that $b(X, K) = q$, then implies $\delta = 0$.

5. We have thus shown that $\Delta(X, K) = 0$ for all X. Condition (9), in combination with our assumption that $b(X_J, J) = \inf_{X' \colon X'_J = X_J} b(X', K)$, then implies $\Delta(X_J, J) = 0$ for all X, J. The claim follows.

We now show the reverse claim, that any extremal point of $\mathcal{B}$ needs to satisfy these conditions. If any of these conditions is violated, we can construct a $\Delta \not\equiv 0$ which satisfies the conditions of Lemma 1.

1. Suppose first that there are two points X, X' such that $0 < q_1 = b(X, K) < b(X', K) = q_2 < 1$, so that the first condition of the proposition is violated. Let $q_0 < q_1 < q_2 < q_3$ be four adjacent points in the range of $b(X, K)$.[8] Denote $p_1 = \mathrm{P}_{\theta_0}(b(X, K) = q_1)$ and $p_2 = \mathrm{P}_{\theta_0}(b(X, K) = q_2)$, and set

$$\epsilon = \min(q_1 - q_0, q_2 - q_1, q_3 - q_2),$$
$$\rho_1 = \begin{cases} 1 & \text{if } p_1 = p_2 = 0 \\ p_2 & \text{else} \end{cases}, \qquad \rho_2 = \begin{cases} 1 & \text{if } p_1 = p_2 = 0 \\ p_1 & \text{else} \end{cases}.$$

Define
$$\Delta(X_J, J) = \begin{cases} \epsilon \cdot \rho_1 & \text{if } b(X_J, J) = q_1 \\ -\epsilon \cdot \rho_2 & \text{if } b(X_J, J) = q_2 \\ 0 & \text{else.} \end{cases}$$

This Δ satisfies the conditions of Lemma 1.

2. Suppose next that the first condition of the proposition holds, and there exists X' such that $0 < b(X', K) = q < 1$, but $\mathrm{P}_{\theta_0}(b(X, K) = q) = 0$, so that the

[8] This is the only point in the proof where we use that $b(X, K)$ has finite range.

second condition of the proposition is violated. Define

$$\Delta(X_J, J) = \begin{cases} \min(q, 1-q) & \text{if } b(X_J, J) = q \\ 0 & \text{else.} \end{cases}$$

This Δ satisfies the conditions of Lemma 1.

3. Suppose lastly that the first two conditions of the proposition hold, but that the third condition of this proposition is violated. In that case there must be two points $X' \neq X''$ such that $b(X', K) = b(X'', K) = q$, and we have that $b(X'_J, J) = 0$ for all J such that $X''_J = X'_J$.

   Denote $p_1 = \mathrm{P}_{\theta_0}(X')$ and $p_2 = \mathrm{P}_{\theta_0}(X'')$, and set

$$\epsilon = \min(q, 1-q),$$
$$\rho_1 = \begin{cases} 1 & \text{if } p_1 = p_2 = 0 \\ p_2 & \text{else} \end{cases}, \qquad \rho_2 = \begin{cases} 1 & \text{if } p_1 = p_2 = 0 \\ p_1 & \text{else} \end{cases}.$$

   Define

$$\Delta(X_J, J) = \begin{cases} \epsilon \cdot \rho_1 & \text{if } J = K, X = X' \\ -\epsilon \cdot \rho_2 & \text{if } J = K, X = X'' \\ 0 & \text{if } J = K, X \neq X', X'' \\ \Delta(X, K) & \text{if } J \neq K, b(X_J, J) = b(X, K) = q \\ 0 & \text{else.} \end{cases}$$

   The penultimate line is well-defined since there is at most one such X (among X' and X'') for any given X_J, J, such that $b(X_J, J) = b(X, K) = q$, given our assumptions. This Δ once again satisfies the conditions of Lemma 1.

□

Proof of Proposition 6:

We construct a prior $\mathrm{P}_\pi(X_J, J)$ such that $\mathrm{P}_\pi(J = K) = 1$, and such that b is optimal within the set of functions b that satisfy size control and the support condition. It then follows that b is also optimal within the smaller set $\mathcal{B}$.

We can define P_π as follows:

$$d\,\mathrm{P}_\pi(X_J, J) = \begin{cases} 0 & \text{if } J \neq K \\ d\,\mathrm{P}_{\theta_0}(X, K) \cdot (2-\alpha) & \text{if } b(X,K) = 1, J = K \\ d\,\mathrm{P}_{\theta_0}(X, K) \cdot (1-\alpha) & \text{if } b(X,K) = 0, J = K \end{cases}$$

By size control, $\mathrm{P}_{\theta_0}(b(X,K) = 1) = \alpha$. This implies that $d\,\mathrm{P}_\pi(X_J, J)$ integrates to 1. Furthermore, a simple Langrangian calculation shows that b is optimal for the problem of maximizing $\int b(X_K, J) d\,\mathrm{P}_\pi(X_J, J)$ subject to the support condition $b \in [0,1]$, and subject to the size constraint. $\square$

B Numerical examples

In this appendix, we discuss some numerical examples of solutions to the analyst's problem, for the case of optimal testing. These examples are based on the code for our interactive app (`https://maxkasy.shinyapps.io/The_PAP_App/`), and demonstrate how the app might be used. These examples illustrate that the conclusions of Proposition 5 indeed hold, and that, subject to the characterizations of Proposition 5, a wide range of tests might be optimal, depending on the parameters of the problem. For each example, we report the *optimal full data test*. The test actually implemented is then based on worst-case assumptions about unreported components of the full data.

Simple tests In the following, we also report the optimal *simple* test. The idea here is to restrict the analyst's choice set at the pre-analysis stage, in the mechanism of Theorem 2, by requiring them to report a test $t \in \mathcal{T}' \subseteq \mathcal{T}$, where $\mathcal{T}$ is the set of full-data tests satisfying size control. More specifically, the simple tests $\mathcal{T}'$ that we consider are cutoff tests of the form $t(X) = \mathbf{1}(\sum_{i \in M} X_i > z) + \kappa \cdot \mathbf{1}(\sum_{i \in M} X_i = z)$, where the index set M is chosen by the analyst, and the cutoff z and the rejection probability at the margin κ are then pinned down by the requirement of size control.

The rationale for such simple tests is that they might be easier to report and interpret, relative to the fully optimal implementable tests. This might come at a cost in expected power, however, as the following examples demonstrate.

For all of our examples, we assume that the availability of components X_i is independent across i, conditional on the analyst's information, and that component X_i is available with probability η_i.

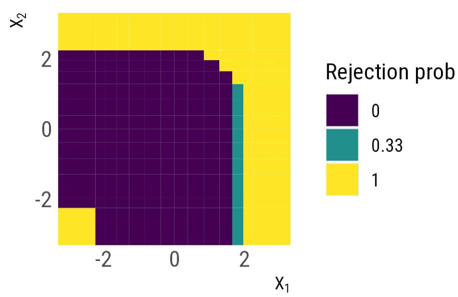
B.1 Normal data

Our first set of examples is of the form considered in Section 2, where the components X_i are normally distributed. The X_i might for instance correspond to the estimated treatment effect for different outcomes of the same treatment, or for different subpopulations. We assume that $X \sim N(\mu_0, \Sigma_0)$ under the null hypothesis. We assume furthermore that $X|\pi \sim N(\mu, \Sigma)$ under the analyst (interim) posterior. Throughout the following examples, the null hypothesis is that $\mu_0 = 0$ and $\Sigma_0 = I$. The required size of the test is .05.

To transform the analyst's problem into a linear programming problem that is numerically tractable, we discretize the support of X, based on the marginal quantiles of the components X_i under the null hypothesis. We then consider full-data tests that are constant within the cells defined by this discretization.

Example 1 The probability of observing each of the components is (0.9, 0.5).
The interim prior is that X has a mean vector of (1, 1), and a variance of $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

OPTIMAL TEST



Expected power: 0.312

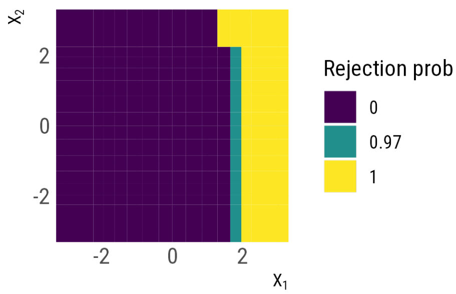
OPTIMAL SIMPLE TEST

Parameter	Value
Components	1
Cutoff for sum	1.645

Expected power: 0.292

Example 2 The probability of observing each of the components is (0.9, 0.1).
The interim prior is that X has a mean vector of (1, 1), and a variance of $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

Optimal test	Optimal simple test



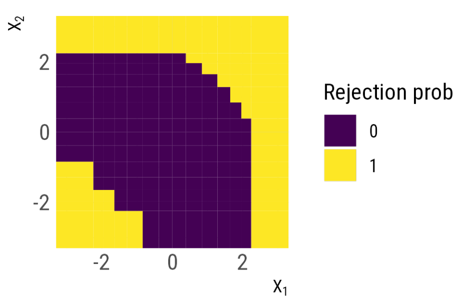
Parameter	Value
Components	1
Cutoff for sum	1.645

Expected power: 0.292

Expected power: 0.292

Example 3 The probability of observing each of the components is (0.9, 0.9). The interim prior is that X has a mean vector of (1, 1), and a variance of $\left(\begin{smallmatrix}3 & 2\\ 2 & 3\end{smallmatrix}\right)$.

Optimal test	Optimal simple test



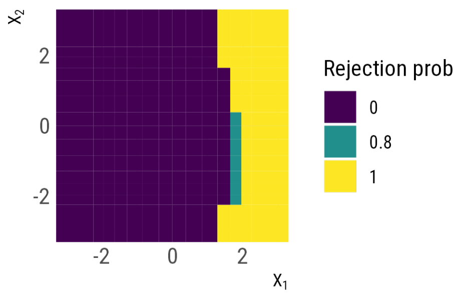
Parameter	Value
Components	1,2
Cutoff for sum	2.326

Expected power: 0.425

Expected power: 0.372

Example 4 The probability of observing each of the components is (0.9, 0.9). The interim prior is that X has a mean vector of (2, 0.5), and a variance of $\left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right)$.

Optimal test	Optimal simple test
Expected power: 0.542	Expected power: 0.539



Parameter	Value
Components	1
Cutoff for sum	1.645

B.2 Binary data

For our second set of examples, we assume that the components X_i are binary, and Bernoulli distributed with expectation θ. The X_i might for instance correspond to the outcome of different tests of the same compound null hypothesis. Throughout the following examples, the null hypothesis is that $\theta \leq .1$. The required size of the test is .05. The assumed (interim) prior for θ is the uniform distribution over $[0, 1]$.

Example 5 The probability of observing each of the components is (0.9, 0.5).

Optimal test

X1	X2	t
1	0	0.44
1	1	1.00

Expected power: 0.283

Optimal simple test

Parameter	Value
Components	1,2
Cutoff for sum	1
Rejection prob at the margin	0.19

Expected power: 0.228

Example 6 The probability of observing each of the components is (0.9, 0.5, 0.1).

Optimal test				Optimal simple test	
X1	X2	X3	t	Parameter	Value
1	0	0	0.44	Components	1,2
1	0	1	0.44	Cutoff for sum	1
1	1	0	1.00	Rejection prob at the margin	0.19
1	1	1	1.00		
Expected power: 0.283				Expected power: 0.228	

Example 7 The probability of observing each of the components is (0.9, 0.8, 0.7, 0.6).

Optimal test					Optimal simple test	
X1	X2	X3	X4	t	Parameter	Value
0	0	1	1	0.72	Components	1,2,3
1	1	0	1	1.00	Cutoff for sum	1
1	1	0	0	1.00	Rejection prob at the margin	0.05
1	0	1	0	1.00		
0	1	1	0	1.00		
1	1	1	0	1.00		
1	0	0	1	1.00		
0	1	0	1	1.00		
1	0	1	1	1.00		
0	1	1	1	1.00		
1	1	1	1	1.00		
Expected power: 0.467					Expected power: 0.4	