Optimal Pre-Analysis Plans: Statistical Decisions Subject to Implementability

Maximilian Kasy Jann Spiess

July 17, 2024

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. One example is drug approval with diverging objectives between a regulator and a pharma company. Another example is hypothesis testing with diverging objectives between readers and authors of research. In our 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 of optimal rules requires pre-analysis plans. Pre-analysis plans allow the decision-maker to leverage analyst expertise. Applying these results to hypothesis testing, 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 Stefano DellaVigna, Ted Miguel, Marco Ottaviani, and Davide Viviano, as well as Alex Frankel, Carlos Gonzalez Perez, Rohit Lamba, Ludvig Sinander, and 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."

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. As a potential solution, 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 can enable valid inference by pre-specifying a mapping from the data to testing decisions or estimates, cf. Christensen and Miguel (2018); Miguel (2021). This can 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 clarify the benefits and optimal design of pre-analysis plans by modeling statistical inference as a mechanism-design problem (Myerson, 1986; Kamenica, 2019). To motivate this approach, note that, in single-agent statistical decision theory, rational decision-makers with preferences that are consistent over time do not need 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. Single decision-makers have no reason to "cheat themselves." The situation is different, however, when there are multiple agents with conflicting interests. When there are multiple agents, not all statistical decision rules might be implementable. Furthermore, allowing for messages (PAPs) before the data are seen can increase the set of implementable rules, and thus improve welfare.[3]

Our framework provides a theoretical justification of PAPs. In addition to our theoretical results, which are based on this framework, we also derive 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, and describe how to derive optimal pre-analysis

[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.

[3] A separate argument for pre-analysis plans, which we do not pursue in this paper, might be based on dynamic inconsistencies in agent preferences, for instance, because of present-bias.

plans from the analysts' perspective. These pre-analysis plans maximize power while controlling size and maintaining implementability. We furthermore provide software (an interactive web app) to facilitate the design of optimal pre-analysis plans.

Examples In our model, we consider the interaction between a decision-maker and an analyst. The analyst has private information and interests which differ from those of the decision-maker. 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 timeline of our model is as follows. Before observing the data, the analyst can send a message to the decision-maker. This message might for instance be in the form of a pre-analysis plan. Then the analyst observes the data. The data are given in the form of a set of statistics, such as the outcomes of different hypothesis tests, or estimates for different model specifications. The analyst chooses a subset of these statistics to report to the decision-maker.

The decision-maker observes the pre-analysis message and the statistics which the analyst reported, and makes a decision based on this information. We assume that this decision is real-valued, and that the analyst always prefers a higher value for this decision. We consider different objectives for the decision-maker, including statistical testing subject to size control.

In our model, the analyst can *hide* information from the decision-maker, by not reporting some statistics, but they cannot *lie* about the data that they report. The potential value of a pre-analysis message in this model comes from the fact that it allows the analyst to share private information (i.e., expertise) with the decision-maker. Sharing such information truthfully would not be incentive-compatible if the

message could only be sent after seeing the data. The analyst might have private information regarding the availability of statistics, and regarding the state of the world.

To make it possible for the analyst to hide information, they need to have plausible deniability: The decision-maker does not know what statistics the analyst got to see. Experiments might not have been run, or data might not have been collected, for instance. The analyst might also have prior uncertainty over the availability of statistics, but this is not necessary for our conclusions.

The mechanism-design approach which motivates our model 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 implementability constraints that are a consequence of the social nature of research.

Implementable decision rules For this model, we first characterize the set of implementable statistical decision rules. This set is independent of decision-maker preferences. We show that implementable decision rules are such that reporting more results can never make the analyst worse off, given the pre-analysis message, and given the realization of the data. Formally, implementable decision rules need to be *monotonic in the reported set* of statistics, in terms of set inclusion.

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).

Pre-analysis messages allow the decision-maker to implement a larger set of decision rules than would be available without such messages. Implementable rules can be implemented using different mechanisms, based on such pre-analysis messages. One possible implementation allows the analyst to *choose from a restricted set of decision rules* before seeing the data. Each of these rules needs to be monotonic in the set of reported statistics. This implementation 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's 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).

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 statistics that the analyst might observe, where this test controls size, and then (ii) implementing this test, making *worst-case assumptions* about any unreported statistics.

The analyst's problem of finding a full-data test that maximizes expected power for this mechanism can be cast as a linear programming problem. If the the analyst knows the set of available statistics at the time of writing their pre-analysis plan, this problem reduces to the classic problem of find a test based on the full set of available statistics with high expected power, subject to size control. The solution to this problem takes the form of a likelihood ratio test. More generally, the set of available statistics might not be known for sure at the time of writing the PAP. We provide an interactive app that allows the analyst to solve the linear programming problem, based on their prior beliefs. The output of our 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. In Section 6, we illustrate our results by applying them to the setting of DellaVigna and Pope (2018), using expert forecasts to construct a prior distribution. In Section 7, we summarize and discuss some limitations of our model. Appendix A contains all proofs.

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), which are motivated by the distortions to statistical inference that might be induced by selective reporting, cf. Andrews and Kasy (2019); Andrews et al. (2023). 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, as a motivating example and special case. The full data consists of two normally distributed statistics, $X = (X_1, X_2)$, with $X_i \sim \mathcal{N}(\theta, 1)$, independently across components of the vector X. The X_i might for instance correspond to experimental estimates of an average treatment effect, for two different experimental sites. There is a decision-maker and an analyst. The decision-maker wants to test the null hypothesis $H_0 : \theta \leq 0$. The analyst, however, aims to simply maximize the probability of rejection.

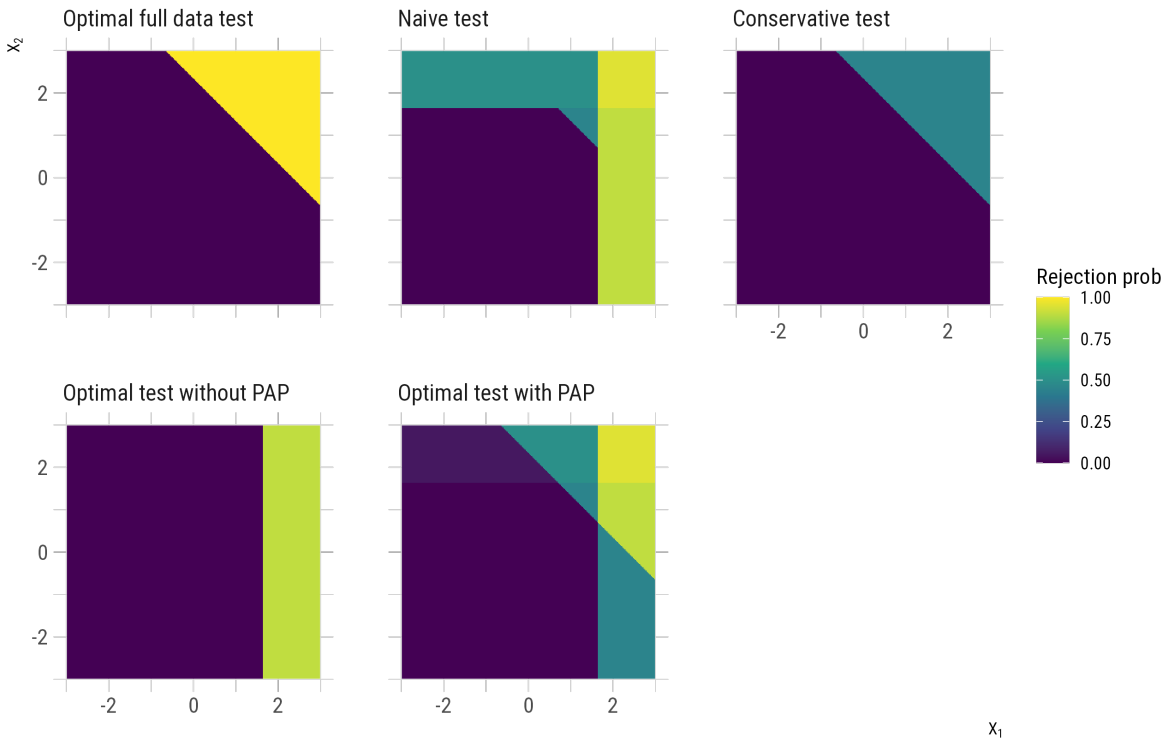
The analyst might not always observe both statistics X_1, X_2. They instead observe the subvector X_J for a random index set J. The possible values of the index set J are $\emptyset$, $\{1\}$, $\{2\}$, and $\{1, 2\}$. The statistic X_i, for $i \in \{1, 2\}$, is observed with probability $P(i \in J)$. Observability is independent across statistics. $P(i \in J)$ is the decision-maker's a-priori probability that the analyst successfully implemented an experiment at site i.

The decision-maker does not know which statistics are actually available, that is, they do not know J. The analyst knows which statistics are available. This allows the analyst to selectively report ("p-hack"), with plausible deniability, since they might not have observed some unreported statistic. Upon learning the data X_J, the analyst chooses a subset $I \subseteq J$, and reports (X_I, I) to the decision-maker. The decision-maker then rejects the null with probability $\mathbf{a}(X_I, I) \in [0, 1]$. 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, $\mathbf{a}_1$ through $\mathbf{a}_5$. For each of these testing rules, Figure 1 shows the rejection probability as a function of (X_1, X_2), assuming that $P(1 \in J) = 0.9$ and $P(2 \in J) = 0.5$. This rejection probability conditions on X, but averages over the distribution of J, and takes into account the analyst's endogenous response to a given testing rule. 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.

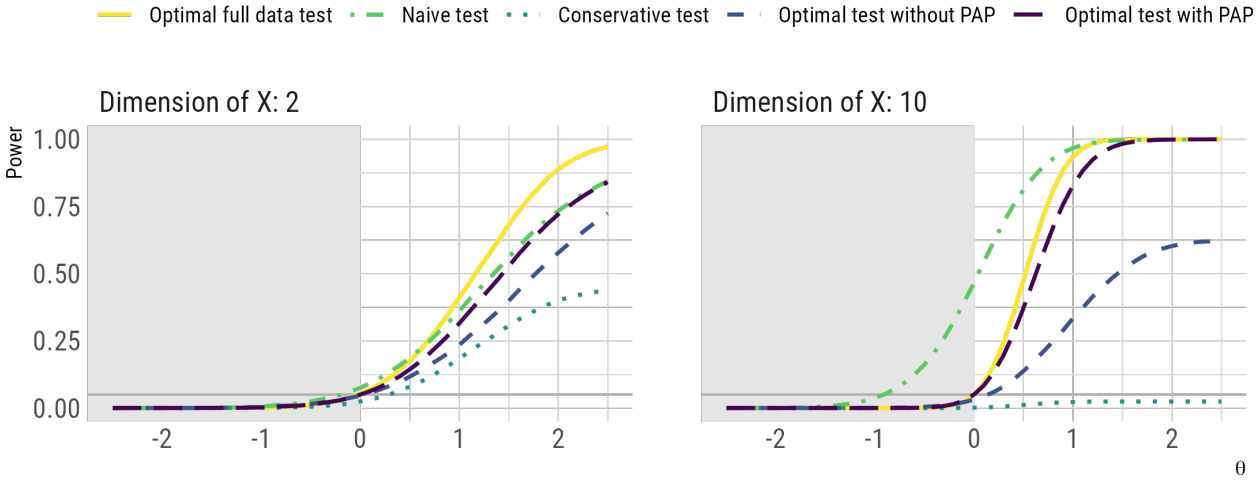
Our benchmark is the **optimal test using all the data**. This test is not, in general, feasible, since not all statistics are always available. We have that $Z =$

Figure 1: Rejection probabilities for different testing rules



$X_1, X_2 \sim N(\theta, 1)$, independently. H0: $\theta < 0$. Probabilities of observing X_1 and X_2 are 0.9 and 0.5.

Figure 2: Power curves



H0: $\theta < 0$. Nominal rejection probability: .05

$\frac{1}{\sqrt{2}}(X_1 + X_2) \sim \mathcal{N}(\sqrt{2} \cdot \theta, 1)$ 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 $\mathbf{a}_1(X) = \mathbf{1}(Z > z)$, 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 statistics I are the full data available to the analyst, and implements the corresponding uniformly most powerful test,

$$\mathbf{a}_2(X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{|I|}} \sum_{i \in I} X_i > z\right).$$

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 the 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 statistics. This results in the **conservative test**,

$$\mathbf{a}_3(X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{2}}(X_1 + X_2) > z \text{ and } I = \{1, 2\}\right).$$

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

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 statistics, 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

$$\mathbf{a}_4(X_I, I) = \mathbf{1}\left(X_1 > z \text{ and } 1 \in I\right).$$

This test implies a probability of rejection given X of $P(1 \in J) \cdot \mathbf{1}\left(X_1 > z\right)$. This

test is optimal for some parameter values, while in general, the optimal test depends on the decision-maker's prior.[4] 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 in this section the analyst knows the statistics J that are available before possibly reporting a PAP, but we assume that they have no private information regarding θ or X. (We relax these assumptions in our general setup below.) The optimal implementable solution can be implemented as follows: The analyst communicates which statistics are available by sending the pre-analysis message $M = J$, and the test is given by

$$\mathbf{a}_5(M, X_I, I) = \mathbf{1}\left(\frac{1}{\sqrt{|M|}} \cdot \sum_{i \in M} X_i > z \text{ and } M \subseteq I\right).$$

That is, the analyst commits to reporting all statistics in J, and for that set of statistics, 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 $n = 10$, where the probability $P(i \in J)$ of observing each of the statistics X_i is evenly distributed over a grid from .5 to .9. The latter case illustrates the contrasts 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 statistics 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 than either of these rules. It controls size, and is in fact strictly conservative under the null. At the same time, it has non-trivial power, which greatly exceeds that of the conservative test. This test without a PAP remains itself far from optimal, however. The optimal

[4]For the given prior over J, this test is for instance optimal when expected power is calculated using the degenerate prior $P(\theta = .3) = 1$. More generally, whether this rule is optimal depends on the prior for both θ and J.

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.

3 Setup

We next describe our general setup, which will be discussed for the rest of this paper. Our setup consists of a game between a decision-maker and an analyst. This game is summarized in Assumption 1.[5] The corresponding timeline is shown in Figure 3. Throughout, X is a collection of statistics X_i, where $i \in \{1, \ldots, n\}$. I and J are (random) index sets, $I, J \subset \{1, \ldots, n\}$, and $X_I = (X_i)_{i \in I}$ denotes the subset of statistics corresponding to the index set I.

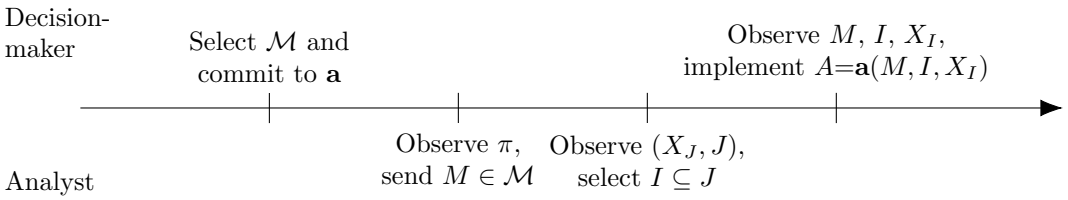
Assumption 1 (Setup)**.** *The game between decision-maker and analyst unfolds as follows:*

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 given θ, J, π only depends on θ, i.e., $X|\theta, J, \pi \stackrel{d}{=} X|\theta$.*

[5] 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.

Figure 3: Timeline



Discussion This is a game of partial verifiability. The report X_I is always truthful given I, but the non-availability of the statistics corresponding to $\{1, \ldots, k\} \setminus J$ cannot be verified by the decision-maker. *Selective reporting*, where not all available statistics are reported ($I \subsetneq J$), corresponds to p-hacking, or specification searching. Misreporting 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 J of statistics X_i. Without such uncertainty, the mechanism design problem would be trivial, and the decision-maker could simply require the analyst to report everything, by threatening to take action $\min \mathcal{A}$ otherwise. Prior uncertainty allows for *"plausible deniability,"* because the decision-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, and more realistically, 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; Assumption 1 remains silent on these. We allow for *conflicting objectives*, which render the mechanism-design problem non-trivial. By contrast, we have already

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 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 decision A is real-valued, i.e., $A \in \mathcal{A} \subset \mathbb{R}$. 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 set J of available statistics 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

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

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.[6]

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:*

[6]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.

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 allowed to choose the decision function from a pre-specified, restricted set $\mathcal{B}$. This second implementation corresponds closely to the actual practice of pre-analysis plans. In this implementation, 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* $\mathcal{B}$*, chosen by the decision-maker, 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 then discuss the connection of truthful revelation to proper scoring. We also consider variants of the model where decision-functions are constrained to be in some class of suitably simple functions.

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 every implementable decision functions $\bar{\mathbf{a}}$ is almost surely identical to a function $\tilde{\mathbf{a}}$ in the convex polytope characterized by the following constraints:

$$\tilde{\mathbf{a}}(\pi, X_J, J) \in \mathcal{A}, \qquad \text{(Support)}$$

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

$$\sum_{X_J, J} (\tilde{\mathbf{a}}(\pi', X_J, J) - \tilde{\mathbf{a}}(\pi, X_J, J))\ \mathrm{P}_\pi(X_J, J) \leq 0 \quad \forall\, \pi', \pi. \qquad \text{(Truthful message)}$$

In the last inequality, P_π is a shorthand for the analyst's posterior distribution conditional on π. This characterization of the implementable set follows immediately from Theorem 1.

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).

Let us elaborate on this equivalence. 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 by Here we assume for simplicity that X has finite support, though the argument generalizes. $\langle f(\cdot), g(\cdot)\rangle = \sum_{X_J,J} f(X_J, J) \cdot g(X_J, J)$, where the running indices X_J, J are understood here as values, rather than random variables. We obtain the following characterization, which was first stated by Savage (1971) and is restated as Theorem 2 in Gneiting and Raftery (2007). Recall that P_{π} is the distribution of (X_J, J) given π.

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 of P_{π}, with sub-gradient G', such that $G(\mathrm{P}_{\pi}) = S(\pi, \pi)$ on the support of π, and 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 chooses and communicates one of the decision functions $b \in \mathcal{B}$ before gaining 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.

One example of such a restricted set $\mathcal{B}_0$ are the index rules implemented in our

app, which is described below. These index rules are of the form

$$b(X_I, I) = \mathbf{1}\left(I \subseteq I_b \text{ and } \sum_{i \in I_b} X_i \geq z_b\right),$$

where I_b is the set of statistics included in the index, and z_b is a critical value.

4.3 Are pre-analysis messages needed?

Aligned objectives Why does implementability in our setting 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 of (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 furthermore 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 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 3 (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* $\bar{\mathbf{a}}$ *which satisfies* $0 \leq \bar{\mathbf{a}} \leq 1$ *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, Assumption 2, and Assumption 3, 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.[7] 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 compound hypotheses. Denote $K = \{1, \ldots, k\}$ the index set of all potentially available statistics. Let $\mathcal{B}$ be the set of measurable functions $b(X_J, J)$ defined by the following constraints.

$$\begin{aligned} &\int b(X, K)\, d\,\mathrm{P}_{\theta_0}(X) \leq \alpha, && && \text{(Size control)} \\ &b(X_J, J) \in [0, 1] && \forall\ J, X, && \text{(Support)} \\ &b(X_J, J) \leq b(X, K) && \forall\ J, X. && \text{(Monotonicity)} \end{aligned} \tag{7}$$

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.

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

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 $b \in \mathcal{B}$.[8]

The case of known J The analyst's problem simplifies to the standard problem of finding a test of maximal expected power subject to size control, if we assume that the analyst knows the value of J, at the time of specifying their PAP. Let J' be this known non-random value of J. Under this assumption, the optimal implementable test is a function of $X_{J'}$ only, and can be written as a likelihood ratio test.

Proposition 5. *Suppose that Assumption 1, Assumption 2, and Assumption 3 hold, and consider the mechanism specified in Theorem 2. Suppose additionally that $P_\pi(J = J') = 1$ for some non-random value J'. Then there exists a solution b to the analyst's problem such that $b(X_K, K) = b(X_{J'}, J')$ for all values of X.*

Any solution of the analyst's problem that is of this form furthermore satisfies that

$$b(X_K, K) = \begin{cases} 1 & \text{when } d\,\mathrm{P}_\pi(X_{J'}, J') > \kappa \cdot d\,\mathrm{P}_{\theta_0}(X_{J'}, J') \\ 0 & \text{when } d\,\mathrm{P}_\pi(X_{J'}, J') < \kappa \cdot d\,\mathrm{P}_{\theta_0}(X_{J'}, J') \end{cases}.$$

for some critical value κ.

Proposition 5 implies that the null should be rejected based on the value of the likelihood ratio test statistic $\frac{d\,\mathrm{P}_\pi(X_{J'}, J')}{d\,\mathrm{P}_{\theta_0}(X_{J'}, J')}$ (assuming this statistic is well defined). Note that the likelihood in the numerator $d\,\mathrm{P}_\pi(X_{J'}, J')$ is in fact the *marginal* likelihood under the interim prior given π, averaging over both the prior for θ, J', and over the sampling distribution of X given θ. See Lehmann and Romano (2006) (Section 3.8) for a discussion of statistical tests that maximize weighted average power, and of least favorable priors.

Potentially optimal tests: Extremal points of $\mathcal{B}$ Let us now return to the more general case, where the analyst does not necessarily know the value of J after

[8] This app is available at https://maxkasy.github.io/home/pap_app.

observing π. Suppose we maintain Assumption 1, Assumption 2, and Assumption 3, 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 6.

- *Suppose that Assumption 1, Assumption 2, and Assumption 3 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 additionally 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 $\mathrm{P}_{\theta_0}(t(X) = q) > 0$.*
  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 6 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 7. *Consider $b \in \mathcal{B}$ with $\mathrm{P}_{\theta_0}(b(X, K) \notin \{0, 1\}) = 0$, and such that the size constraint is binding. 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.

6 Case study

We next discuss a numerical example, to illustrate our results on optimal pre-analysis plans for hypothesis testing. Our example is calibrated to the data and the priors reported in DellaVigna and Pope (2018), who experimentally evaluate 15 different treatments to induce costly effort, in addition to 3 control treatments. The outcome X_i is the effect of treatment i on the average number of button presses, in an Amazon Mechanical Turk task. We consider the effect relative to a control treatment, where participantes are paid 1 cent per 100 button presses. DellaVigna and Pope (2018) also report prior predicted treatment effects, as elicited from 208 academic experts.

We use these expert predictions to calibrate our prior for $\theta = (\theta_i)_{1\leq i\leq 15}$, where θ_i is the true effect of treatment i. We assume that $\theta \sim N(\mu, \Sigma)$ is jointly normal, with prior mean μ equal to the averages of expert forecasts, and prior variance Σ equal to the variance across forecasts. We furthermore assume that the estimated treatment effects have a sampling distribution of $X_i \sim \mathcal{N}(\theta_i, \sigma_i^2/n + \sigma_0^2/n)$, where the sample size is $n = 100$, and σ_0^2/n is the variance of the mean outcome for the control treatment.[9] The standard deviations σ_i are known and correspond to the standard errors reported in DellaVigna and Pope (2018).

We lastly assume, for the purpose of illustration, that the analyst only intends to run experiments for two of the 15 experimental treatments, corresponding to arm 1 (4 cents per 100 presses), and arm 2 (a lottery with a chance of winning 1 dollar per 100 presses with 1% probability). We assume, for now, that the analyst knows ex-ante that $J = \{1, 2\}$. Consider the null (joint) null hypothesis that there are no treatment effects for any of the incentive schemes, $\theta_i = 0$ for all i.

The optimal PAP What is the optimal PAP for this null? The answer is given by Proposition 5. The optimal PAP, for known J, pre-specifies a test which rejects whenever both components of J are reported, and $\log\left(\frac{d\mathrm{P}_\pi(X_J,J)}{d\mathrm{P}_{\theta_0}(X_J,J)}\right)$ exceeds some

[9] The variation across experts is different from the variance of the prior of any individual expert. We furthermore deviate the original sample size of around 550 per arm in the paper. For these reasons, or numerical example should only be thought of as a calibration for the purpose of illustrating our theory.

critical value. Under our assumptions,

$$\log\left(\frac{dP_\pi(X_J,J)}{dP_{\theta_0}(X_J,J)}\right) = const. + \left(\begin{smallmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{smallmatrix}\right)' S^{-1} \left(\begin{smallmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{smallmatrix}\right) - \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)' S_0^{-1} \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)$$

where S_0 is the sampling variance of X_J, and S is the prior variance of X_J, which equals the sum of the prior variance of θ_J plus the sampling variance,

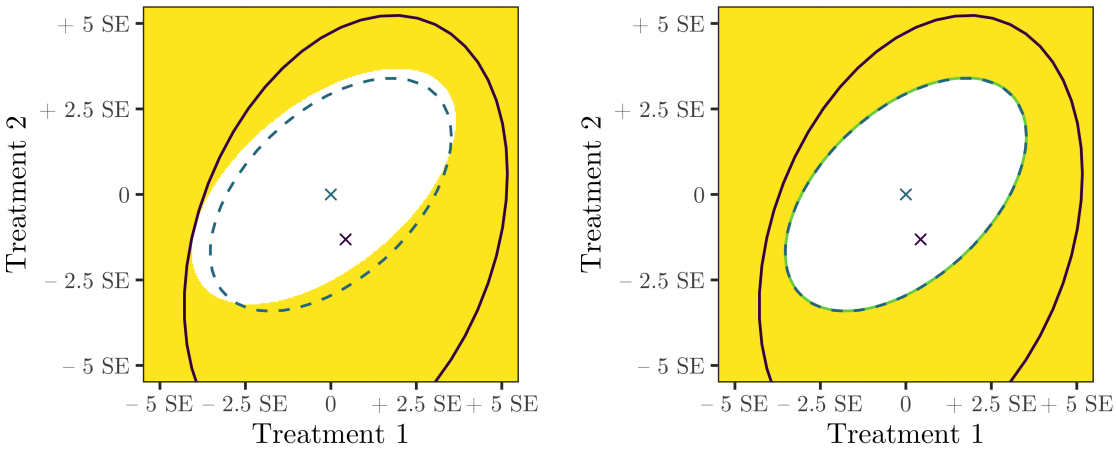
$$S_0 = \frac{1}{n}\left(\begin{smallmatrix} \sigma_0^2+\sigma_1^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2+\sigma_2^2 \end{smallmatrix}\right), \qquad\qquad S = \left(\begin{smallmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{smallmatrix}\right) + S_0.$$

We visualize this test in Figure 4a. The axes of the graph represent estimated treatment effects, normalized by their sampling standard error, $X_i/\sqrt{(\sigma_0^2+\sigma_i^2)/n}$, for $i = 1, 2$. The blue ellipse (dashed line) represents the null distribution of X_J, with 95% of draws falling within the circle; and the purple ellipse (solid line) represents the prior marginal distribution of X_J. The optimal rejection region at a 5% size is shaded in yellow. The likelihood-ratio test of Proposition 5 yields an ellipsoidal rejection region.

An optimal simple PAP In practice, fully optimal tests may be hard to describe in a PAP. What is the optimal PAP subject to an additional simplicity constraint? Let us restrict attention to tests that reject if the test statistic $X_J' \cdot S_0^{-1} \cdot X_J$ exceeds some critical value c_J and if all components in J are reported, where both J and the critical value are pre-specified. This is the standard Wald (χ^2) test for the subset J. Subject to this restriction, it is optimal for the analyst to pre-register their true $J = \{1, 2\}$, and a critical value of 6 (for a test of size .05). We visualize this test in Figure 4a. Restricting tests to be simpler leads to a loss in average power, but this loss is small in our numerical example. Average power of the optimal test is approximately .52. The restriction to a Wald test reduces power to .50.

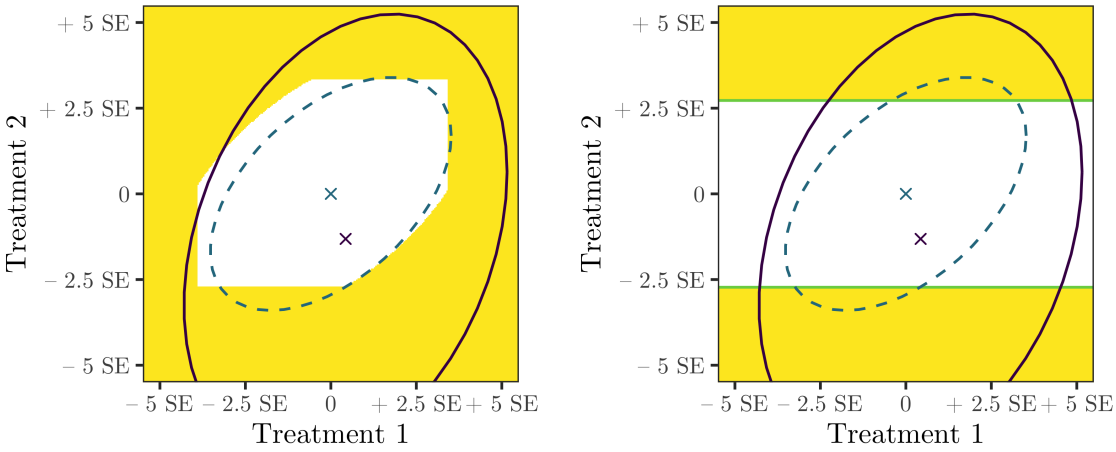
Analyst uncertainty about J Assume now that the analyst is uncertain about which components J will be available. Maybe some experiments are not always feasible, or data collected differ from those in the original plan. Assume that arm 1 is available with ex-ante probability .5, and arm 2 with probability .7, independently across arms. In this case, the rejection regions of the optimal PAP are more complex, and are given by the solution to the linear programming problem discussed in Section 5.3. Figure 5a plots the optimal test, which solves this linear program. If

Figure 4: Analyst knows J



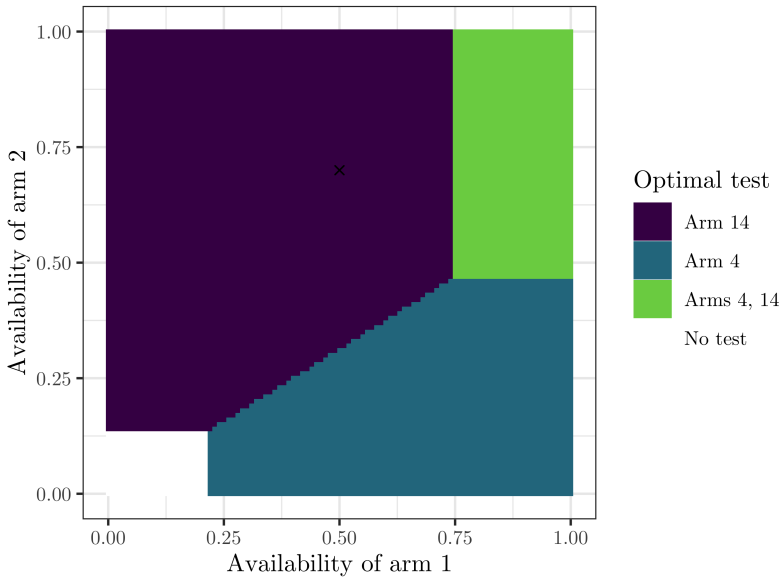
(a) Optimal PAP (b) Optimal simple PAP

Figure 5: Analyst is uncertain about J



(a) Optimal PAP (b) Optimal simple PAP

Figure 6: Which treatment arms should be registered?



one of the arms is not available in the end, then the decision-maker makes worst-case assumptions about this arm, and implements the corresponding testing decision. Because the components i are not always available, overall expected power only equals .32 in this example.

We can also, again, consider simple PAPs, which specify Wald tests for some pre-selected set of components J'. The optimal PAP of this form ignores arm 1 and specifies a standard two-sided t-test that rejects for $\frac{|X_2|}{\sqrt{\sigma_2^2+\sigma_0^2}} > 1.96$ (Figure 5b). That is, despite arm 1 being available some of the time, it is better to only consider arm 2 in this case. This result is driven by the different priors over the effect of these treatments, as well as by different availabilities, where arm 2 is more likely to lead to a rejection and is more likely to be available. Restricting attention to such a simple test reduces expected power from .32 (for the optimal test) to only .26.

Concluding our discussion of this numerical example, we plot in Figure 6 how the optimal set of pre-registered components J', for a simple test, depends on the probability that data for either treatment is available. Appendix B elaborates further.

7 Conclusion

We conclude by summarizing our main contributions, before discussing some limitations of our model and avenues for future research. We have proposed a principal-agent model of pre-specification in empirical research. In our model, a decision-maker relies on the examination and reporting of data by an analyst. The analyst can selectively report statistics that they observe, but they cannot lie about the observed statistics. The decision-maker does not know which data are available to the analyst. This allows for plausible deniability.

Our model provides a theoretical justification for pre-analysis plans (or, more generally, pre-analysis messages), which cannot be rationalized in traditional single-agent statistical decision theory. There is no need for sending messages prior to seeing data in the single agent framework - in fact, there would not even be a recipient for such a message in this framework.

The constraint of implementability in our model leads to a constrained version of statistical decision-theory. Constrained optimal decision functions generally require a PAP. PAPs allow the decision-maker to draw on analyst expertise. Such analyst expertise cannot be used under the alternative of unilateral specification of decision functions by the decision-maker.

Our model also allows us to derive practical guidance for the design of optimal PAPs. Optimal PAPs lead to constrained-optimal decision functions. We show that the decision-maker's optimal decision function can be implemented by allowing the analyst to choose from a restricted set of decision-functions, and communicating their choice in a PAP. For hypothesis testing, the analyst gets to choose any test which satisfies size control when all data are observed. If a statistic required by the pre-specified test is not reported, then the decision-maker later makes worst-case assumptions about this statistic. The analyst problem, for this mechanism, reduces to a linear programming problem. They have to maximize expected power subject to size control, and subject to the constraints implied by implementability. When the set of available statistics is known to the analyst in advance, then the solution to the analyst problem takes the form of a likelihood ratio test. More generally, we provide an app which allows the analyst to easily solve their optimization problem.

Our model is quite general in describing the problem of selective reporting by an analyst with conflicting objectives and private expertise. There are some important

considerations, however, which are not reflected in this model, for the sake of analytical clarity. First, we do not model the potential cost to researchers of documenting complex estimation and testing procedures in the PAP. This is a cost which has been emphasized by critics of the widespread adoption of PAPs (Coffman and Niederle, 2015; Olken, 2015; Duflo et al., 2020). Relatedly, we do not model the costs of communicating complex findings. Such costs likely play an important role in explaining why not all findings are published (Frankel and Kasy, 2022; Andrews and Shapiro, 2021).

Second, there are a number of alternative mechanisms which might complement PAPs as tools to limit the adverse effects of conflicting interests and private information. One such mechanism is adversarial review, where reviewers might request additional statistics to be reported by researchers. Our model does not include a review stage. Another such mechanism is researcher reputation, and more generally the dynamics of repeated interactions. Our model is a one-shot game, which does not allow for such dynamics. We hope that future research will elaborate on mechanisms such as these, and the extent to which they might act as a substitute for PAPs.

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*.

DellaVigna, Stefano and Devin Pope (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

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. □

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 $\mathcal{M} = \mathcal{B}$. □

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. □

Proof of Proposition 3.
Denote by

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

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 $\tilde{\mathbf{a}}$. □

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 a best-response

of the analyst 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. □

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. □

Proof of Proposition 5:
Let $\tilde{b}$ be some solution of the analyst's problem Define

$$b(X_J, J) = \begin{cases} \tilde{b}(X_{J'}, J') & J' \subseteq J \\ 0 & \text{else.} \end{cases}$$

Then $0 \leq b(X_J, J) \leq \tilde{b}(X_J, J)$ for all X and J, and b satisfies all the constraints if $\tilde{b}$ does. Furthermore, expected power for b is the same as expected power of $\tilde{b}$, since the two functions are identical on the support of P_π. Therefore b is a solution of the analyst's problem.

The second claim follows from an application of the Neyman–Pearson Lemma (cf. Theorem 3.2.1 in Lehmann and Romano 2006) to the point null hypothesis P_{θ_0} and the point alternative P_π. □

To prove Proposition 6, 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:*

$$\int \Delta(X, K) d\,\mathrm{P}_{\theta_0}(X) = 0 \tag{8}$$

$$|\Delta(X_J, J)| \leq \min(b(X_J, J), 1 - b(X_J, J)) \qquad \forall\, J, X \tag{9}$$

$$|\Delta(X_J, J) - \Delta(X, K)| \leq b(X, K) - b(X_J, J) \qquad \forall\, J, X. \tag{10}$$

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

Proof of Proposition 6:
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 (9), $\Delta(X, K) = 0$ 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 (10) 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 (8), 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 (10), in combination with our assumption that $b(X_J, J) = \inf_{X' :\, X'_J = X_J} b(X', K)$, 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 $b(X, K)$.[10] Denote

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

$p_1 = \mathrm{P}_{\theta_0}(b(X,K) = q_1)$ 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 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 7:

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 assumption size control is binding, $\mathrm{P}_{\theta_0}(b(X, K) = 1) = \alpha$. This implies that $d\,\mathrm{P}_\pi(X_J, J)$ integrates to 1. Furthermore, a simple Lagrangian 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. □

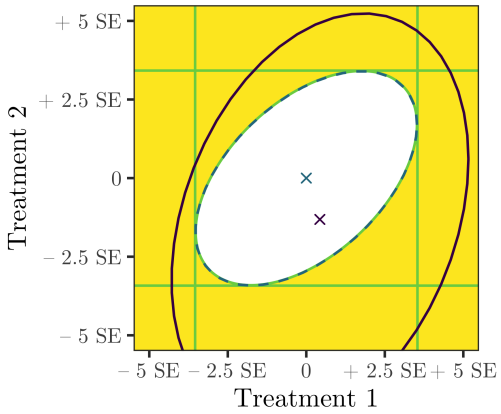
B Case study continued: Simple PAPs

In Section 6, we reported first-best optimal PAPs, as well as optimal simple PAPs that can be represented as Wald tests for pre-specified subsets I. In this section, we consider the alternative restriction to PAPs where the set of components I that will be submitted does not have to be pre-specified, but the analyst needs to pre-specify different thresholds c_I for Wald tests for different sets I. Optimal simple tests of this form are reported in Figure 7. When the analyst knows that arms 1 and 2 are available, the optimal such test is the same as the simple test from Figure 4b. If, however, the analyst is uncertain about which arms will be available, then the optimal subset-specific thresholds are non-trivial. The resulting test in Figure 7b rejects if either the t-statistic for arm 1 exceeds 3.64, or the t-statistic for arm 2 exceeds 2.92, or the Wald statistic for arms 1 and 2 exceeds 2.85[2]

This test approximates the optimal test from Figure 5a well, with only a small loss of power from 32% to 31% due to the simplicity restriction. Reporting such data-specific thresholds may represent a practical way of committing to effective tests without communicating overly complex rejection regions.

Figure 7: Optimal simple pre-specified rejection regions for arm-specific cutoffs

(a) Known availability



(b) Uncertain availability