(When) should you adjust inferences for multiple hypothesis testing?*

Davide Viviano
UC San Diego

Kaspar Wüthrich
UC San Diego

Paul Niehaus
UC San Diego

July 27, 2021

Abstract

The use of multiple hypothesis testing adjustments varies widely in applied economic research, without consensus on when and how it should be done. We provide a game-theoretic foundation for this practice. Adjustments are often — but not always — appropriate in our framework when research influences multiple *policy decisions*. These adjustments depend on the nature of scale economies in the research production function and on economic interactions between policy decisions, with control of classical notions of compound error rates emerging in some but not all cases. When research examines multiple *outcomes*, on the other hand, this motivates either very conservative testing procedures or aggregating outcomes into sufficient statistics for policy-making.

Keywords: Bonferroni, family-wise error rate, false discovery rate, multiple subgroups, multiple treatments, multiple outcomes, research costs, interaction effects

*Viviano and Wüthrich contributed equally to this work. We are grateful to Nageeb Ali, Pat Kline, Michal Kolesar, Mikkel Plagborg-Møller, Ulrich Mueller, Andres Santos, Azeem Shaikh, Jesse Shapiro, Joel Sobel, Yixiao Sun, and seminar participants at Princeton and UCSD for valuable comments. Muhammad Karim provided excellent research assistance. Wüthrich gratefully acknowledges funding from the UCSD Academic Senate. All errors and omissions are our own. Email: {dviviano, kwuthrich, pniehaus}@ucsd.edu

1 Introduction

Empirical papers in economics usually test more than one hypothesis. Historically researchers have treated these tests as independent when conducting inference. But recently some have begun to approach "multiple hypothesis testing" (MHT) scenarios differently, using procedures that adjust one test for the presence of others in order to control an aggregate error rate (e.g. the family-wise error rate (FWER) or the false discovery rate (FDR)). There is now wide variation both in the use of such adjusted procedures and in the specific procedures used. For example, 41% of experimental papers published in "top 5" journals in 2020 adopted some form of aggregate error rate control while the other 59% did not.[1] This raises several simple but important questions: when (if at all), how, and why should researchers be required to adjust for multiplicity? While there is a general sense that simply ignoring MHT issues could create problematic incentives, there is little consensus on how exactly they should be addressed. This leaves referees and editors to rely on intuition or taste when evaluating submitted papers.

In this paper we seek to understand whether and why MHT procedures arise as desirable solutions within the research publication process, and if they do, what observable features of the economic environment determine the right procedures to use. To this end we adopt a view of the publication process as a game between a journal editor and a representative researcher, explicitly modelling their incentives and constraints. In particular, the framework embeds three core ideas. First, policy decisions are influenced by the summary recommendations (in particular, hypothesis tests) contained in research papers. Second, while this makes research results a public good, the costs of producing them are born privately by the researcher. She decides whether or not to incur these costs, conducting an experiment with a given number of hypotheses based on its chances of subsequent publication. Anticipating this, the editor must select a hypothesis testing protocol carefully to balance the twin goals of (i) motivating the production of research and (ii) producing good policy guidance from it. Finally, the editor balances these objectives conservatively, selecting protocols that maximize worst-case social welfare.

In the case of a single hypothesis these assumptions can be used to rationalize standard testing procedures, as shown in an insightful paper by Tetenov (2016).[2] To study the multiple hypothesis case, however, we require notions of optimality that are more nuanced than those

[1] Authors' calculations based on a review of papers published in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

[2] While we focus on the research publication process, our framework can also be used to study MHT in other contexts such as regulatory approval as analyzed in Tetenov (2016).

in the single-hypothesis case. We show that with multiple hypotheses no maximin rule Pareto dominates all others (in contrast to the single-hypothesis case). Motivated by this non-existence result, we develop an appropriate notion of *local power* with the property that any maximin rule that is locally most powerful is also admissible, allowing us to select among admissible maximin rules.

A defining feature of our approach is that hypothesis tests correspond to policy decisions; this is what allows us to select between alternative testing procedures based on their welfare consequences. Given this, multiple testing issues arise whenever research informs multiple policy decisions. We begin our analysis with settings in which this happens because research examines multiple *interventions* or effects on multiple *sub-populations*. The mapping between tests and decisions is clear in these cases, and both are common in practice.[3] We focus initially on the case in which hypothesis tests are independent in terms of their economic consequences and their influence on the researcher's publication prospects. For this base case we draw two broad conclusions.

First, it is often optimal to adjust hypothesis testing procedures for the number of hypotheses being tested. A loose intuition for this result is as follows. The worst states of the world from the editor's point of view are those in which the status quo of no intervention is best, as in these states a research study has only downside. In these states the editor maximizes welfare by keeping publication probabilities low enough that the researcher chooses not to experiment. If the hypothesis testing rule were invariant to the number of hypotheses being tested, then for sufficiently many hypotheses this condition would be violated: the chance of getting a study published due solely to false positives would be high enough that the researcher would choose to conduct one. Some adjustment for hypothesis count is thus optimal. We believe that this logic aligns fairly closely with the lay intuition that researchers should not be allowed to test many hypotheses and then "get credit" for false discoveries.

Second (and as this suggests), the research cost function determines whether and how much adjustment is required. We can in fact pick a research cost function such that *no* further adjustment is required, as the costs of doing research scale with the number of hypotheses tested in just such a way as to "build in" the needed correction. A simple example of this occurs when the costs of research increase linearly in the number of hypotheses; in this case no MHT adjustment is required. More generally, the editor selects a testing rule to compensate for residual imbalances in researcher incentives with respect to the number of hypotheses,

[3]For example, 27 of 124 field experiments published in "top-5" journals between 2007 and 2017 feature factorial designs with more than one treatment (Muralidharan et al., 2020). Moreover, the average number of subgroups analyzed in the 34 field experiments published in top economics journals between 2005 and 2009 is 6.4 (Fink et al., 2014).

taking the researcher's costs into account. As a result the framework can explain when common criteria such as control of the FWER and FDR emerge as appropriate solutions for a given economic environment, and when they do not. When research costs are fixed, for example, it is optimal to control the average size of tests (e.g. via a Bonferroni correction), while when costs scale in proportion to the number of tests *no* MHT adjustment is required.

In addition to research costs, we consider a series of extensions in which there is successively more scope for interactions of other kinds. We allow for additional interactions in the rule the journal uses to select publications, examining the effects of a *threshold* rule where papers are published if they find enough results. We then introduce economic interactions in the effects of the treatments being evaluated — allowing for example for complementary interventions. In these more general settings, maximin optimality can be very conservative; we therefore focus on a weaker notion of maximin optimality, corresponding to weak size control in the literature. As in our base model, the broad conclusion remains that some MHT adjustment is warranted to the extent that it is not already "priced in" to the research cost function. Interestingly, we also obtain a form of FWER control in one case, but at the level of *groups* of hypotheses sufficient for publication, and generally not at the level of the individual hypothesis.

Finally, we consider research that examines multiple *outcomes*. If these all inform a single policy decision then this does not lead naturally to MHT adjustments within our framework, as there is in effect only one decision-relevant hypothesis to test. In a paper that examines the effects of an education reform on multiple learning outcomes, for example, the relevant null is that the status quo is better than the reform when those outcomes are evaluated using the policy-maker's preference. In cases like these the editor should require the researcher to test for effects on a function of the outcomes that represents those preferences.[4] This approach is related to the common practice of creating index variables using weights determined by statistical properties of the outcomes (for example, using principal components or inverse (co)variance weighting (e.g., Anderson, 2008)), with the important difference that an outcome's weight should instead reflect its economic importance.

Multiple outcomes may warrant MHT adjustments, on the other hand, if we interpret them as informing an audience of multiple policy-makers with heterogeneous preferences over outcomes. For example, some education policy-makers may care more than others

[4]In an extension we also consider the consequences of uncertainty about the weights the policy-maker places on various outcomes. Under the (pessimistic) assumption that these are chosen adversarially by "nature" this can rationalize testing procedures that vary with the number of hypotheses, but with rather severe penalties for multiplicity, with the optimal level of size control shrinking exponentially with the number of outcomes in simple examples.

about proficiency in math relative to writing. If the editor is unsure which policy-maker will act on the results of the paper and (as before) maximizes worst-case welfare, then she selects hypothesis testing rules that adjust for the number of outcomes. In fact, these adjustments can lead to very conservative procedures such as testing based on the minimum of a group of test statistics using conservative thresholds (or, if the researcher reports separate t-tests, to tests with zero power).

Our paper builds on and is most closely related to work on optimal statistical approaches in economic models, taking into account preferences and incentives (e.g., Chassang et al., 2012; Tetenov, 2016; Spiess, 2018; Henry and Ottaviani, 2019; Yoder, 2019; Banerjee et al., 2020; Williams, 2021). The existing literature does not discuss MHT, which is the focus of our paper. We discuss some practical implications for applied work in Section 5 below.

Our paper also relates to an extensive literature at the intersection between decision theory and hypothesis testing, dating back to Wald (1950) and Robbins (1951). Previous work has motivated notions of compound error control in single-agent non-strategic environments; see, for example, Storey et al. (2003) and Efron et al. (2008) for a Bayesian interpretation of the FDR and Lehmann and Romano (2005b) for a discussion of the FWER. We complement this literature by developing a game-theoretic model of the publication process that explicitly incorporates the incentives and constraints of the different agents involved in the research process. Relative to the decision-theoretic approach, this has two main advantages. First, it lets us characterize *when* MHT adjustments are appropriate — and also when they are *not* — as a function of measurable features of the research and publication process. Second, it allows us to justify and discriminate between different notions of compound error (e.g., FWER and FDR) in the same framework based on these same economic fundamentals.

There is also an extensive statistical literature on MHT focusing on the design of algorithmic procedures for controlling particular notions of compound error.[5] We refer to Efron (2008) and Romano et al. (2010) for overviews. While there are many different MHT procedures, only few statistical optimality results exist (e.g., Lehmann et al., 2005; Romano et al., 2011). We complement these statistical results with a framework that relates the choice of optimal MHT procedures to the economic environment and allows for comparative statics. Finally, we draw on List et al. (2019)'s helpful distinction between different types of multiplicity and show how these distinctions lead to meaningful differences in optimal testing procedures.

On a broader level, our paper is also connected to the literature on statistical treatment

[5]See, e.g., Holm (1979); Westfall and Young (1993); Benjamini and Hochberg (1995); Benjamini and Liu (1999); Storey (2002); Storey et al. (2004); Lehmann and Romano (2005a); Lee and Shaikh (2014); Romano and Wolf (2016); List et al. (2019) among many others.

choice (e.g., Manski, 2004; Manski and Tetenov, 2007; Hirano and Porter, 2009; Tetenov, 2012; Kitagawa and Tetenov, 2018; Hirano and Porter, 2020; Athey and Wager, 2021), which mostly focuses on non-strategic planners' problems, and models of scientific communication (e.g., Frankel and Kasy, 2018; Andrews and Shapiro, 2020).

2 Multiple interventions and multiple subgroups

We model the choice of a statistical testing procedure as part of a game between a representative researcher and a journal editor, building on Tetenov (2016). In our model, multiple testing issues arise whenever research informs multiple policy decisions. We therefore start by discussing settings with multiple *interventions* or different *sub-populations*, as here there is a clear one-to-one mapping between multiple hypothesis tests and multiple policy decisions. The case of multiple *outcomes* is more subtle. In particular, there may not be a one-to-one mapping between hypothesis tests and policy decisions since multiple tests may only inform one decision. We will turn to this case in Section 3.

2.1 Setup and model

The editor prescribes a hypothesis testing protocol, restricting how the researcher can report discoveries and make recommendations. Given this protocol, the researcher decides whether to run an experiment with $J \geq 1$ different treatments. Treatments may represent either different interventions or different sub-populations to whom a given intervention might be applied. What is important is that there are J distinct and non-exclusive policy decisions to be made.

If the researcher experiments, she draws a vector of statistics $X \sim F_\theta$, and incurs a cost $C(J) > 0$, which may depend on the number of treatments. Here, $X \in \mathcal{X} \subseteq \mathbb{R}^J$ and $\theta \in \Theta$ is the parameter of interest. The costs $C(J)$ are sunk after the experiment is conducted and do not depend on θ. We assume that every experiment is written up and submitted to the journal. Note that research designs (defined by J and F_θ) arise exogenously. This captures situations where, for example, researchers collaborate with implementation partners such as NGOs who present them with the opportunity to work on an evaluation whose parameters are largely fixed by the partner's capacity or the size of the population it serves. Research designs are endogenous in the broad sense that the editor's choice of a hypothesis testing protocol determines which designs are implemented and published.

The researcher reports results in the form of a vector of *discoveries* or *recommendations*

$$r(X) = (r_1(X), \ldots, r_J(X))^\top,$$

where $r_j : \mathcal{X} \mapsto \{0,1\}$ and $r_j(X) = 1$ if and only if treatment j is recommended. We will refer to r as *recommendation function* or *hypothesis testing protocol*. We assume that treatments are non-exclusive. The editor chooses the types of statistical test(s) that the researcher may employ by selecting $r \in \mathcal{R}$, where $\mathcal{R}$ is a pre-specified and exogenous class of functions. Unless otherwise specified, we do not impose any restrictions on $\mathcal{R}$ other than pointwise measurability. Our focus will be on understanding how editor-preferred recommendation function(s) vary as a function of variation in the number J of treatments being tested.

We consider a benevolent editor whose utility depends on social welfare, which in turn depends on the researcher's recommendations. Specifically, we assume that the recommended (combination) of treatments is implemented in a target population by a policy-maker who is otherwise a passive player in the game. We assume that the target population is independent of the experimental sample but subject to the same data-generating process.[6] The policy-maker implements the researcher's recommendations irrespective of whether the paper gets published. If on the other hand the researcher does not experiment, the status quo (no treatment) is implemented.

To help simplify our specification of welfare, we introduce the *selector* function δ that indicates which of the $2^J - 1$ possible *combinations* of treatments is recommended,

$$\delta(r(X)) \in \{0,1\}^{2^J-1}, \quad \text{where} \quad \sum_k \delta_k(r(X)) \in \{0,1\}. \tag{1}$$

If no treatment is implemented (i.e. $\sum_k \delta_k(r(X)) = 0$) then the status quo is maintained. For $k = 1, \ldots, 2^J - 1$, let $u_k(\theta)$ denote the social welfare generated by the combination of treatments $\delta_k(r(X))$. We write the welfare effect of implementing the treatments recommended by the researcher concisely as $u(\theta)^\top \delta(r(X))$, where $u(\theta) = (u_1(\theta), \ldots, u_{2^J-1}(\theta))^\top$.

To derive our main results, we assume that welfare is additive.

Assumption 1 (Additive welfare)**.** Suppose that, for $k = 1, \ldots, J$, $u_k(\theta) = \theta_k$, and $u_{J+1}(\theta) = \theta_1 + \theta_2$, $u_{J+2}(\theta) = \theta_1 + \theta_3$, $\ldots$, $u_{2^J-1}(\theta) = \sum_{j=1}^J \theta_j$.

Under Assumption 1, the welfare gains to implementing a combination of treatments is equal the sum of the welfare gains from implementing them individually. This is the case, for example, when the treatments are very different so that interaction effects are unlikely, or when each treatment corresponds to treating a different sub-population and there are no cross-group spillovers. In Section 4, we extend our analysis to general welfare functions that allow for interaction effects.

[6] These are standard assumptions (e.g., Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021).

We will often return to the following running example (or variants of it) in which the researcher studies $J = 2$ interventions using a linear regression model.

Example 1 (Running example)**.** Consider the problem of studying the effect of $J = 2$ non-exclusive experimental treatments D_1 and D_2 on an outcome of interest Y based on a sample with N observations. Suppose that

$$Y_i = \theta_1 D_{i,1} + \theta_2 D_{i,2} + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where σ^2 is known.[7] For simplicity, the baseline average outcome is normalized to zero and θ_1 and θ_2 are the average treatment effects of D_1 and D_2 net of the costs of implementation. Under these assumptions, $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma)$, where $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2\right)^\top$ is the OLS estimator of $\theta = (\theta_1, \theta_2)^\top$ and the covariance matrix Σ is known. In this example, we set $X = \hat{\theta}$. Note that the distribution F_θ depends on the sample sizes in the experimental design, which we treat as exogenous.

The recommendation $r(X)$ can take four values: $r(X) = (0, 0)$ (recommend baseline), $r(X) = (1, 0)$ (recommend D_1), $r(X) = (0, 1)$ (recommend D_2), and $r(X) = (1, 1)$ (recommend D_1 and D_2). The selector δ is defined as $\delta((0, 0)) = (0, 0, 0)$, $\delta((1, 0)) = (1, 0, 0)$, $\delta((0, 1)) = (0, 1, 0)$, and $\delta((1, 1)) = (0, 0, 1)$. A typical choice of the recommendation function (which will be shown to be optimal under some conditions) is

$$r(X) = \left(1\left\{\frac{\hat{\theta}_1}{\sqrt{\Sigma_{1,1}}} \geq t\right\}, 1\left\{\frac{\hat{\theta}_2}{\sqrt{\Sigma_{2,2}}} \geq t\right\}\right)^\top, \tag{3}$$

where the threshold t is chosen (optimally) by the editor. That is, the recommendations are based on standard (one-sided) t-tests. □

We introduce two asymmetries in our model which are essential for justifying hypothesis testing, i.e., a protocol r that imposes a size control criterion. First, following Tetenov (2016) and Di Tillio et al. (2017), we impose asymmetry in the agents' incentives: while the editor maximizes welfare, the experimenter's utility depends not on welfare but on her discoveries. If instead the incentives of the experimenter and the editor were aligned then hypothesis testing of any kind would be unnecessary.[8] Second, we assume, for the moment, that there is asymmetric information: the parameter θ is known to the experimenter but unknown to

[7] This model is similar to the one studied in Section 5.5 of Elliott et al. (2015). The exact normality assumption is imposed for simplicity. Similar results follow from standard asymptotic approximations provided that N is large enough.

[8] Namely, the optimal editor's strategy would be not to impose any restriction on r. In this case the researcher would always report as true discoveries the ones corresponding to positive parameter's values.

the editor. This assumption reflects the fact that researchers typically know more about the specific treatments they study than editors. The editor then uses the hypothesis testing protocol to screen out researchers evaluating treatments that do not benefit social welfare. We relax this assumption in Section A.1, where we allow researchers to only have noisy information on the parameters θ.

The (expected) experimenter's utility conditional on conducting the experiment depends on the costs of doing so and on the (exogenous) probability that this results in a publication, which in turn depends on the underlying parameter θ. The experimenter decides whether or not to experiment based on the expected value $\beta_r(\theta)$ of doing so. We first assume that $\beta_r(\theta)$ is linear in the number of discoveries, and then examine in Section 4 an extension to threshold-crossing publication rules.

Assumption 2 (Linear publication rule)**.** The experimenter's utility conditional on experimenting is, up-to-rescaling by J,

$$\beta_r(\theta) = \int \sum_{j=1}^{J} r_j(x) dF_\theta(x) - C(J), \quad C(J) \leq J. \tag{4}$$

In Assumption 2, we interpret $\int \sum_{j=1}^{J} r_j(x) dF_\theta(x)$ as the publication probability, multiplied by a factor J. Thus, this assumption describes a setting where the publication probability is linear in the number of discoveries. The condition that $C(J) \leq J$ implies that there is at least one case (when all discoveries are reported with probability one) under which it is profitable for the researcher to experiment. Without such an upper bound, the researcher would never experiment.

The following tie-breaking assumption simplifies our analysis by avoiding multiple equilibria.[9]

Assumption 3 (Tie-breaking rule)**.** Whenever the researcher is indifferent between experimenting or not, she makes the choice that yields the highest social welfare.

The editor moves first in the game, and we can analyze her choice taking into account the best response of the experimenter. The editor's utility is given by

$$v_r(\theta) = \begin{cases} \int u(\theta)^\top \delta(r(x)) dF_\theta(x) & \text{if } \beta_r(\theta) > 0, \\ \max\left\{ \int u(\theta)^\top \delta(r(x)) dF_\theta(x), 0 \right\} & \text{if } \beta_r(\theta) = 0, \\ 0 & \text{if } \beta_r(\theta) < 0. \end{cases} \tag{5}$$

The second case ($\beta_r(\theta) = 0$) follows from Assumption 3.

[9] A similar assumption is imposed in Kamenica and Gentzkow (2011).

Note that the structure of the game we study naturally justifies one-sided hypothesis testing, as researchers test whether or not a proposed treatment improves upon the status quo. In Appendix A.2 we outline ways to change the structure of the game in order to justify two-sided hypothesis testing.

We now study hypotheses testing protocols as an optimal strategy of the editor. We first review the single hypothesis case ($J = 1$) previously analyzed by Tetenov (2016) and then generalize to the case with multiple hypotheses.

2.2 Review and illustration with a single hypothesis

Tetenov (2016) considers a game between an informed agent and a regulator, which is relevant, for instance, in the context of drug approvals. We explain his results using the terminology of our game between a researcher and an editor. Without loss of generality, suppose that $u(\theta) \in [-1, 1]$. Define the *null space* of parameter values as the set of parameters such that implementing the (single) treatment being studied would reduce welfare, $\Theta_0 := \{\theta : u(\theta) < 0\}$. Similarly, define the *alternative space* of parameter values as the set of parameters such that the treatment increases welfare $\Theta_1 := \{\theta : u(\theta) \geq 0\}$. Social welfare is $v_r(\theta) = u(\theta)$ if $\int r(x)dF_\theta(x) \geq C$, where we write $C := C(1)$ for simplicity, and zero otherwise. That is, welfare is non-zero if the expected utility from experimenting, $\int r(x)dF_\theta(x)$, is larger than the cost of experimentation.

To justify single hypothesis testing, Tetenov (2016) focuses on maximin optimal recommendation functions, i.e., recommendation functions that maximize worst-case welfare,

$$r^* \in \arg\max_{r \in \mathcal{R}} \min_{\theta \in \Theta} v_r(\theta).$$

The focus on maximin recommendation functions is important for justifying standard hypothesis testing. Intuitively, the worst-case nature of the maximin criterion naturally induces the editor to treat size control and power asymmetrically. Standard hypothesis testing will not generally be optimal under alternative optimality criteria.

Proposition 1 in Tetenov (2016) demonstrates that a recommendation function is maximin optimal if and only if

$$\int r^*(x)dF_\theta(x) \leq C \quad \text{for all} \quad \theta \in \Theta_0. \tag{6}$$

This result shows that maximin optimal recommendation functions are such that the researcher does not find it worthwhile experimenting whenever the treatment is welfare-reducing ($\theta \in \Theta_0$). For this to hold, the probability of publication in this state (given by the left-hand side of (6)) must be sufficiently low. The model thus rationalizes error

control, i.e. control of the probability of falsely rejecting the null that the status quo of no treatment is best.

To select among the many alternative maximin recommendation functions, Tetenov (2016) provides admissibility results under an additional monotone likelihood ratio property. He shows that admissible recommendation functions satisfy the following condition

$$\int r^*(x)dF_0(x) = C, \tag{7}$$

with the recommendation function taking the form of a threshold crossing rule. This result provides a formal justification for standard (one-sided) tests with conventional critical values.

2.3 Maximin recommendation functions and size control

In this section, we discuss the problem in the general case with $J > 1$ treatments. We have seen in Section 2.2 that the editor maximizing worst-case welfare is important for justifying single-hypothesis testing. Since analyzing multiple testing requires a framework for justifying hypothesis testing in the first place, we will focus on maximin optimal recommendation functions; that is, recommendation functions that maximize worst-case welfare[10],

$$r^* \in \arg\max_{r\in\mathcal{R}} \min_{\theta\in\Theta} v_r(\theta).$$

We denote the set of maximin recommendation functions by $\mathcal{M}$.

Define the *null space*, the set of parameters such that welfare is weakly negative regardless of the choice of r, as follows.

Definition 1 (Null space)**.** The null space is $\Theta_0 := \Big\{\theta : u_j(\theta) < 0 \text{ for all } j\Big\}$. □

The following proposition provides a characterization of maximin hypothesis testing protocols r^*, generalizing Proposition 1 in Tetenov (2016) to the case of $J > 1$ hypotheses. It shows that the definition of null space that we employ is directly connected to maximin optimality.

Proposition 1 (Maximin protocols)**.** *Let Assumption 3 hold and suppose that $\Theta_0 \neq \emptyset$. A recommendation function r^* is maximin-optimal, i.e.,*

$$r^* \in \arg\max_{r\in\mathcal{R}} \min_{\theta\in\Theta} v_r(\theta), \tag{8}$$

[10]Maximin hypothesis testing protocols are attractive in settings where there are concerns about researchers experimenting with interventions that may hurt (groups of) individuals. To this end, we will show below that maximin protocols discourage experimentation when treatments have negative welfare impacts. Moreover, focusing on maximin rules allows us to obtain concrete recommendations that do not depend on the editor's prior about θ, which is difficult to model and elicit in practice.

if and only if

$$\beta_{r^*}(\theta) \leq 0 \quad \forall \theta \in \Theta_0 \qquad \textit{and} \qquad v_{r^*}(\tilde{\theta}) \geq 0 \quad \forall \tilde{\theta} \in \Theta \setminus \Theta_0. \tag{9}$$

Proof. See Appendix B.2.1. □

Proposition 1 shows that maximin optimality is equivalent to two conditions. First, as in the case with $J = 1$ hypotheses, maximin recommendation functions depend on the experimenter utility $\beta_r(\theta)$, and deter experimentation over Θ_0, where *all* treatments reduce welfare. Second, the editor's utility for $\theta \in \Theta \setminus \Theta_0$ must be non-negative. This second condition requires that if some treatments reduce the welfare, there must be other treatments that compensate them. The first condition captures notions of size control. We show below that the second condition is non-binding in the leading case where X is normally distributed. Note that Proposition 1 applies very generally, without relying on any particular functional form assumption on the experimenter (or editor) utility.

We illustrate the definition of the null space and the characterization of maximin protocols in our running example.

Example 2 (Running example continued)**.** In our running example, the null space is $\Theta_0 = \{\theta \in \Theta : \theta_1 < 0 \text{ and } \theta_2 < 0\}$. Figure 1 provides a graphical illustration. By Proposition 1, a recommendation function $r^* = (r_1^*, r_2^*)^\top$ is maximin only if (but not necessarily if)

$$P(r_1^*(X) = 1|\theta_1, \theta_2) + P(r_2^*(X) = 1|\theta_1, \theta_2) \leq C(2), \quad \theta_1 < 0, \theta_2 < 0. \tag{10}$$

Equation 10 shows that maxmin recommendation functions impose restrictions on size control (i.e., the probability of reporting a false discovery). □

Remark 1 (Null space)**.** The definition of the null space Θ_0 corresponds to the *global null hypothesis* in the literature. It is a subset of the *strong null space* $\tilde{\Theta}_0 = \{\theta : \theta_j < 0 \text{ for some } j\}$. We note that $\tilde{\Theta}_0$ plays an important role in the second condition of Proposition 1 ($v_r(\tilde{\theta}) \geq 0 \; \forall \tilde{\theta} \in \Theta \setminus \Theta_0$). Since $v_r(\tilde{\theta}) \geq 0$ for all $\tilde{\theta} \in \{\theta : \theta_j \geq 0 \text{ for all } j\}$ by definition, this condition is equivalent to assuming that the editor utility is positive for $\tilde{\theta} \in \tilde{\Theta}_0 \setminus \Theta_0$. □

2.4 Admissibility and power

The set of maximin recommendation functions contains infinitely many elements, some of which may be very conservative. An example is the function that forces the researcher not to report any discoveries ($r_j(X) = 0$ for all j). This motivates the use of additional criteria

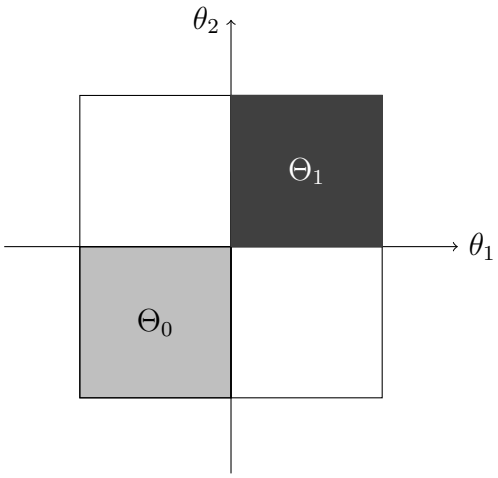


Figure 1: Graphical illustration of the null space $\Theta_0 = \{\theta \in \Theta : \theta_1 < 0 \text{ and } \theta_2 < 0\}$ and the alternative space $\Theta_1 = \{\theta \in \Theta : \theta_1 \geq 0 \text{ and } \theta_2 \geq 0\}$. See Remark 1 for a discussion of the two orthants where the coefficients have different signs.

for choosing between them. Among recommendation functions that minimize the editor's downside, in other words, how might she select those with large upside?

We proceed as follows. First, we show that no maximin recommendation function Pareto dominates all other recommendation functions when $J > 1$. This is in sharp contrast to the single-hypothesis case in Section 2.2, where Pareto dominant recommendation functions exist. Second, motivated by this result, we introduce a notion of local power and show that locally most powerful recommendation functions are admissible. Finally, we provide explicit characterizations of locally most powerful recommendation functions.

We start by defining a suitable notion of weak Pareto dominance.

Definition 2 (Weak Pareto dominance[11])**.** The recommendation function r *weakly* Pareto dominates r' if $v_r(\theta) \geq v_{r'}(\theta)$ for all $\theta \in \Theta$. A Pareto dominant recommendation function r over a set $\tilde{\mathcal{R}} \subseteq \mathcal{R}$ is a recommendation function that *weakly* Pareto dominates all other recommendation functions $r' \in \tilde{\mathcal{R}}$. □

Recall that maximin recommendation functions discourage experimentation, leading to zero editor utility for $\theta \in \Theta_0$. Therefore, Pareto dominance among such rules is determined by their relative performance in $\Theta \setminus \Theta_0$ and, thus, captures a notion of power in our setting.

To derive the remainder of our results, we focus on the leading case where X is normally distributed.

[11] This definition is expressed in terms of weak inequality in order to accommodate sets of different functions r having the same probabilities of discoveries.

Assumption 4 (Normality)**.** The statistic X is normally distributed, $X \sim \mathcal{N}(\theta, \Sigma)$, where Σ is positive definite.

The following proposition shows that there exists a data-generating process such that no maximin recommendation function weakly Pareto dominates all other recommendation functions. This result holds in particular when X is normally distributed.

Proposition 2 (No recommendation function Pareto dominates the others)**.** *Let $J > 1$. Let Assumptions 1, 2, and 3 hold. Then there exists a parameter space $\Theta \subseteq [-1, 1]^J$ and a distribution $\{F_\theta, \theta \in \Theta\}$ such that no maximin recommendation function r (weakly) Pareto dominates all other maximin recommendation functions $r' \in \mathcal{M}$ for any cost $0 < C(J) \leq (J-1)$. Moreover, there exists such a distribution F_θ that satisfies Assumption 4.*

Proof. The proof of Proposition 2 is based on the following observation. For any maximin recommendation function we can find a θ with one (or more) positive component such that the recommendation function is dominated by another function with the property that the positive component has the largest probability of discovery, and the remaining entries have zero probability of discovery. See Appendix B.2.2 for details. □

In the terminology of classical hypothesis testing, Proposition 2 states that there are settings in which no uniformly most powerful test exists (where in our setting power is measured in terms of implied welfare).[12] It implies that we cannot find recommendation functions that are Pareto dominant for all alternatives.

Motivated by this result, we focus instead on ranking maximin recommendation functions within a particular set of *local* alternatives to the null space Θ_0.

Definition 3 (ϵ-alternatives)**.** For $\epsilon > 0$, define the local alternative space as

$$\Theta_1(\epsilon) := \Big\{\theta : u_j(\theta) \geq \epsilon \text{ for some } j, u_j(\theta) \geq 0 \text{ for all } j\Big\}.$$

□

The set of ϵ-alternatives $\Theta_1(\epsilon)$ is the set of parameters for which, for some decision, the editor's utility is strictly positive by at least ϵ. Note that $\Theta_1(\epsilon) \cap \Theta_0 = \emptyset$ for all $\epsilon \geq 0$.

Based on Definition 3, we introduce the following notion of local power.

[12] This result differs from single-hypothesis testing, where under normality the most powerful test exist by classical results when interpreting power in terms of welfare effects (see, for example, Chapter 15 in Van der Vaart, 2000).

Definition 4 (ϵ-more powerful)**.** A recommendation function r is ϵ-more powerful than r' if[13]

$$\liminf_{\epsilon \downarrow 0} \left\{ \frac{1}{\epsilon} \inf_{\theta \in \Theta_1(\epsilon)} v_r(\theta) - \frac{1}{\epsilon} \inf_{\theta' \in \Theta_1(\epsilon)} v_{r'}(\theta') \right\} > 0.$$

□

Definition 4 introduces a partial ordering of recommendation functions based on their worst-case performance under ϵ-alternatives. It considers parameter values in an alternative space that contains the origin as $\epsilon \to 0$. The difference of the utilities is rescaled by the location parameter ϵ.[14]

We say that a maximin recommendation function r is *ϵ-most powerful* if it is ϵ-more powerful than any other maximin recommendation function r'. The following proposition confirms that any maximin ϵ-most powerful recommendation function is also *admissible* (i.e., not strictly Pareto dominated by any other recommendation function).

Proposition 3 (Admissibility)**.** *Suppose that $u_j(\theta) \in [-b, b]$ for a positive constant $b > 0$. Then any maximin and ϵ-most powerful recommendation function r is admissible.*

Proof. See Appendix B.2.3. □

Proposition 3 motivates our focus on maximin and ϵ-most powerful recommendation functions in the following sections.[15]

Remark 2 (Trivial maximin decision rules are not admissible)**.** Suppose that there exists a recommendation function r that is maximin and ϵ-most powerful. Then maximin recommendation functions that are not ϵ-most powerful are not guaranteed to be admissible. For example, the function $r(X) = (0, \ldots, 0)^\top$ is maximin optimal. However, it is not admissible if a non-trivial maximin decision rule exists (see, for example, Proposition 4). □

2.5 Most powerful maximin functions

This section provides explicit characterizations and examples of ϵ-most powerful and maximin recommendation functions with an additive welfare (Assumption 1) and a linear publication

[13] In Appendix B.2.3, we show that the expression below is uniformly bounded for all (r, r').

[14] Rescaling by the location parameter is common in local asymptotic analyses and is standard practice when making optimality statements (e.g., Athey and Wager, 2021). Without rescaling the difference would trivially converge to zero.

[15] The notion of admissibility is treating as inadmissible any recommendation function that is dominated by any other recommendation function (including recommendation functions that are non maximin optimal). Note, however, that maximin recommendation functions, by definition, cannot be dominated by non-maximin decisions over Θ_0.

rule (Assumption 2). In Section 4 we extend our analysis to general welfare functions and alternative publication rules.

The following lemma provides a characterization of the ϵ-most powerful recommendation functions.

Lemma 1 (Separate size control is ϵ-most powerful)**.** *Let $J > 1$. Let Assumptions 1, 2, 3, and 4 hold, and let $\Theta = [-1, 1]^J$. Then $r^* \in \mathcal{R}$ is maximin optimal and ϵ-most powerful if and only if r^* satisfies Equation (9) and*

$$\lim_{\theta \downarrow 0} P\Big(r_j^*(X) = 1|\theta\Big) = P(r_j^*(X) = 1|\theta = 0) = \frac{C(J)}{J} \quad \forall j \in \{1, \ldots, J\}, \tag{11}$$

assuming that such r^* exists.

Proof. See Appendix B.1.1. We note that the proof does not rely on normality of X (Assumption 4). We only require that X is continuously distributed with CDF F_θ, which admits a PDF $f_\theta(x)$ that is continuous in θ for all $x \in \mathcal{X}$. □

Lemma 1 has two important implications. First, it shows that a recommendation function is ϵ-most powerful and maximin if and only if it imposes size control that is *separate*, meaning that given restrictions on the marginal probabilities of rejecting individual hypotheses no further restrictions are placed on the joint probabilities. Second, Lemma 1 shows that whether and to what extent the level of these separate tests should depend on the number of hypotheses being tested — in other words, whether an adjustment for the presence of multiple hypothesis is required — depends on the structure of the research production function $C(J)$. In particular, if costs scale linearly with J then no adjustment is required, while if there are economies of scale in the research production function ($C(J)$ scales sub-linearly with J) then some adjustment is required. In the extreme case, where $C(J)$ is fixed, $C(J) = \alpha$, Lemma 1 implies that classical Bonferroni corrections are optimal:

$$P(r_j^*(X) = 1|\theta = 0) = \frac{\alpha}{J} \quad \forall j \in \{1, \ldots, J\}.$$

We provide a further discussion of the role of the cost function in Section 5.

An important question is what types of recommendation functions r^* are maximin and ϵ-most powerful. The next proposition shows that under normality the widely-used *threshold-crossing* recommendation functions corresponding to standard one-sided tests are maximin optimal and ϵ-most powerful.

Proposition 4 (Optimality of separate t-tests)**.** *Assume that the conditions in Lemma 1 hold and $C(J) > 0$. Then the recommendation function*

$$r_j^*(X) = 1\Big\{X_j/\sqrt{\Sigma_{j,j}} \geq t\Big\}, \quad \forall j \in \{1, \ldots, J\} \tag{12}$$

is maximin optimal and ϵ-most powerful, for $t = \Phi^{-1}(1 - C(J)/J)$.

Proof. See Appendix B.2.4. □

Proposition 4 shows that a standard one-sided t-test with critical value $\Phi^{-1}(1 - C(J)/J)$ is optimal in our setting. The critical value depends on the number of hypotheses J and the structure of the cost function $C(J)$.

Remark 3 (Average size control is optimal)**.** Lemma 1 shows that most powerful maximin rules control average size, $\sum_{j=1}^{J} P(r_j^*(X) = 1|\theta = 0) = C(J)$. Many of the popular MHT corrections reviewed in the introduction do not directly target average size control and, thus, will generally not be optimal in our model.

This explains why classical Bonferroni corrections are optimal in our model when $C(J)$ is constant, while common refinements of Bonferroni such as Holm (1979)'s method are not. By construction, Bonferroni satisfies average size control, whereas common refinements do not. The optimality of Bonferroni (and average size control) is driven by our choice of the publication rule. In Section 4, we will show that Bonferroni corrections may not be optimal with other publication rules. □

3 Multiple outcomes

In the framework above we interpret a hypothesis test as a recommendation about a policy decision. Papers that examine multiple *interventions* or impacts within multiple *sub-groups* could clearly inform multiple policy decisions — which of the interventions to implement, or which of the sub-groups to treat. In some cases this gives rise as we have seen to a rationale for MHT adjustments.

With multiple *outcomes*, there may not be such a one-to-one mapping between hypothesis tests and policy decisions. This is because a paper that examines multiple outcomes may or may not inform multiple policy decisions. For example, a paper that measures the impact of a single reform on multiple measures of well-being might be used to guide the single decision whether or not to scale up that reform. In this case our framework interprets the paper as making a single recommendation, and thus does not rationalize MHT adjustments. On the other hand, one might interpret a paper that examines multiple outcomes as informing multiple decisions by different policy-makers, some of whom care more about some outcomes than others. In this case we can interpret the paper as making multiple recommendations, and the question whether to apply an MHT adjustment to those recommendations is well-posed.

In this section we formalize these ideas and, in the latter case of a heterogeneous audience, characterize the forms of MHT adjustment the editor prefers. We examine the case of a single treatment in order to focus attention on issues that are specific to multiple outcomes; the results can be extended to multiple treatments at the expense of additional notation.

3.1 Single policy-maker

Consider first as a benchmark case an audience of a single policy-maker considering a single intervention whose payoffs from implementing it are a function of G different outcomes, $(Y_1, \ldots, Y_G)$. Associated with the G outcomes is a vector of welfare functions $(u^1(\theta), \ldots, u^G(\theta)) = (\theta_1, \ldots, \theta_G)$ whose entries correspond to the social benefits (or costs) of the treatment due to its effects on those outcomes. Overall welfare is a weighted average of these functions,

$$u_{w^*}(\theta) = \sum_{g=1}^{G} w_g^* u^g(\theta), \tag{13}$$

For example, the policy-maker might care about the impacts of a microfinance intervention on both livelihoods and on measures of women's empowerment.

We assume here that the weights $w^* = (w_1^*, \ldots, w_G^*)^\top$ (and thus $u_{w^*}(\theta)$) are common knowledge. Indeed, one implication of our framework is that it is important to the publication process for the researcher to elicit them so that his paper can report a test of the appropriate index. This contrasts with indexing procedures often used in practice in which researchers aggregate multiple outcomes using weights that depend on their statistical properties (e.g. inverse variance weighting or principal component analysis), as opposed to their economic significance. In Appendix A.4 we consider the consequences of uncertainty about the policy-relevant weights and show that — if the editor applies worst-case logic as above — this leads to extremely conservative testing procedures.

For tests of the known index $u_{w^*}(\theta)$ the results in Tetenov (2016) and our previous results directly apply with $u(\theta) = u_{w^*}(\theta)$ and

$$v_r(\theta; w^*) = \begin{cases} \int r(x) u_{w^*}(\theta) F_\theta(x) & \text{if } \beta_r(\theta) > 0 \\ \max\left\{ \int r(x) u_{w^*}(\theta) F_\theta(x), 0 \right\} & \text{if } \beta_r(\theta) = 0 \\ 0 & \text{if } \beta_r(\theta) < 0. \end{cases}$$

MHT adjustments in the usual sense do not arise here because there is only a single hypothesis of interest, $u_{w^*}(\theta) < 0$, and there is ultimately only a single policy decision to be made. For example, if the outcomes Y_g are different components of expenditure but the policy-maker cares only about increasing total expenditure $Y = \sum_g Y_g$ then there is no need to test for

effects on the components individually, and hence no room for MHT adjustments in the usual sense.[16]

3.2 Multiple policy-makers

An alternative interpretation of the multiple outcomes setting is that the audience for the research includes multiple policy-makers with heterogeneous valuations of the outcomes. For example, the education ministry in one country may be more concerned about literacy, while the education ministry in another neighboring country places relatively more weight on quantitative skills. To capture this idea, consider a setup in which the audience for the research includes G different policy-makers. We assume without loss of generality that policy-maker g's welfare effect corresponds to the effect on outcome Y_g, captured by $u^g(\theta) = \theta_g$ (for example, Y_g might be a weighted average of an underlying set of outcomes using weights that reflect g's preferences). The experimenter makes G recommendations, one for each policy-maker such that recommendation functions take the form

$$r : \mathcal{X} \mapsto \{0,1\}^G, \quad (r_1(X), \ldots, r_G(X))^\top,$$

where $r_g(X)$ is the recommendation to policy-maker g.

As in the previous section, policy-maker g acts as a passive players in our game. She implements the policy recommended by the editor based on the gth outcome. As a result, policy-maker g's utility conditional on experimentation is $P(r_g(X) = 1|\theta)u^g(\theta)$. The experimenter's expected utility $\beta_r^y(\theta)$ depends on the publication prospects and the costs; we will consider concrete examples below. We use the superscript y to distinguish the objects in this section from those introduced before because the model is different.

We assume the editor is aware of the different policy-makers who may read and implement the papers' recommendations, but does not know for certain *which* will do so. He thus faces two sources of uncertainty, with respect to both the welfare effects of the treatments and the audience for evidence on those effects. Extending the approach above, we assume the editor chooses a testing protocol to maximize worst-case welfare over both θ and g, the identity of the implementing policy-maker. Therefore, the editor's utility can be written in compact form as

$$v_r^y(\theta) = \begin{cases} \min_{g\in\{1,\ldots,G\}} P(r_g(X) = 1|\theta)\theta_g & \text{if } \beta_r(\theta) > 0 \\ \max\left\{ \min_{g\in\{1,\ldots,G\}} P(r_g(X) = 1|\theta)\theta_g, 0 \right\} & \text{if } \beta_r(\theta) = 0 \\ 0 & \text{if } \beta_r(\theta) < 0. \end{cases}$$

[16] In Appendix A.4, we consider the consequences of uncertainty about the weights and show that such uncertainty can rationalize testing procedures that vary with the number of hypotheses.

The null space takes the following form:

$$\Theta_0^y := \{\theta : u^g(\theta) < 0, \text{ for some } g \in \{1, \ldots, G\}\} \tag{14}$$

Intuitively, worst-case welfare is negative if any of the $u^g(\theta)$ is negative. To make explicit the connection to the null space with a single outcome in Section 2.2, note that

$$\{\theta : u^g(\theta) < 0, \text{ for some } g \in \{1, \ldots, G\}\} = \cup_{g=1}^{G} \{\theta : u^g(\theta) < 0\}, \tag{15}$$

where each component in the right-hand side denotes the null space with a single outcome and a single hypothesis. This shows that the null space with multiple outcomes can be very large when there are many outcomes, which we will show leads to conservative hypothesis testing protocols.

The next proposition characterizes maximin protocols with multiple policy-makers.

Proposition 5 (Multiple policy-makers)**.** *Let Assumption 3 hold and suppose that* $\Theta_0^y \neq \emptyset$*. A recommendation function* r^* *is maximin-optimal with multiple outcomes, i.e.,*

$$r^* \in \arg\max_{r \in \mathcal{R}} \min_{\theta \in \Theta} v_r^y(\theta)$$

if and only if

$$\beta_{r^*}^y(\theta) \leq 0 \quad \forall \theta \in \Theta_0^y, \tag{16}$$

where Θ_0^y *is defined in Equation* (14)*.*

Proof. See Appendix B.2.9. □

It is interesting to compare the result in Proposition 5 to that with multiple treatments in Proposition 1 because, in both settings, the research reports a vector of discoveries. The key difference is that the null space with multiple outcomes, Θ_0^y, contains not only parameter values for which all components are negative but those for which *some* components are negative.[17] As a result we will see that this setup rationalizes stricter notions of size control than above, which can be conservative.

To see why Proposition 5 can lead to conservative hypothesis testing protocols, consider a linear publication rule as in Assumption 2,

$$\beta_r^y(\theta) = \int \sum_{g=1}^{G} r_g(x) dF_\theta(x) - C(G), \tag{17}$$

[17] Using standard hypothesis testing terminology, the null space corresponds to strong error control; see the discussion in Remark 1 and Lehmann and Romano (2005b).

for some cost function $C(G)$. Proposition 5 implies that if we restrict the class of feasible recommendation function to the (common) threshold crossing rules, power needs to be zero for the rule to be maximin-optimal. Specifically, suppose that $\beta_r^y(\theta)$ is as in Equation (17) and that $X \sim \mathcal{N}\Big(\theta_g, I\Big)$, where $\theta_1, \ldots, \theta_G \in [-M, M]$, for some arbitrary large M. Consider the following threshold crossing recommendation function

$$r(X) = \left(1\left\{X_1 \geq t\right\}, \ldots, 1\left\{X_G \geq t\right\}\right)^\top,$$

and let $C(G) = \alpha G$ for some $\alpha \in (0, 1)$. Then for M large enough and $\alpha \leq 1 - 1/G$, the threshold crossing rule is maximin optimal only in the trivial case where $t \to \infty$, which implies that the tests never reject and have zero power.[18]

A natural question is how non-trivial maximin recommendation functions look like in this setting. For this, we need to move away from threshold-crossing rules. One example of a maximin recommendation function is

$$r_g(X) = 1\left\{\min_{g \in \{1,\ldots,G\}} X_g \geq t\right\}, \quad \forall g \in \{1, \ldots, G\}. \tag{18}$$

where the threshold t is chosen such that $P(X_g \geq t | \theta_g = 0) = C(G)/G$ for all g.[19] Due to the use of the minimum across all statistics and the choice of the threshold, the recommendation function (18) can be very conservative when there are many outcomes.

An important aspect of Proposition 5 is that it allows us to directly map the features of the publication process and researcher utility to different types of compound error rate control. Consider, for instance, the case of a threshold crossing publication rule, where papers are published if they "have done enough", i.e., if they report more than κ discoveries

$$\beta_r^y(\theta) = \gamma \int 1\left\{\sum_{g=1}^{G} r_g(x) \geq \kappa\right\} dF_\theta(x) - C(G). \tag{19}$$

Here the probability of publication is equal to γ if the number of discoveries exceeds κ, so that the paper has "done enough" to be considered for publication, and zero otherwise. We assume that $\gamma \in (C(G), 1)$ such that there is a least one case where experimentation is profitable; otherwise the researcher would never experiment. With a threshold crossing publication rule the incremental value to the researcher of rejecting any given hypothesis may depend on the number of other hypotheses also rejected.

[18] To see this, take $\theta_g \to \infty$ for all $g < G$ and $\theta_G < 0$. Then $\sum_{g=1}^{G} P(r_g(X) = 1|\theta_g) = P(r_g(X) = 1|\theta_G) + G - 1$ such that we need to impose that $P(r_g(X) = 1|\theta_G) \leq \alpha G - G + 1$, where $\alpha G - G + 1 \leq 0$ for $\alpha \leq 1 - 1/G$.

[19] To see why this decision rule is maximin, note that $\sum_g P(r_g(X) = 1|\theta_g) = G \prod_{g=1}^{G} P(X_g \geq t|\theta_g)$. For some $g' \in \{1, \ldots, G\}$, let $\theta^g \to \infty$ for all $g \neq g'$, and $\theta_{g'} \leq 0$. Then we obtain that the expression is bounded from above by $GP(X_{g'} \geq t|\theta_{g'} = 0) = C(G)$.

Then Proposition 5 implies that r^* is maximin if and only if

$$P\left(\sum_{g=1}^{G} r_g(X) \geq \kappa | \theta\right) = P\left(\text{at least } \kappa \text{ discoveries} \mid \theta\right) \leq \frac{C(G)}{\gamma} \quad \forall \theta \in \Theta_0^y$$

This criterion is quite restrictive; it is stronger than and implies strong control of the κ-FWER at level $C(G)/\gamma$. To illustrate, let $\kappa = 1$. Then we impose restrictions not only on the probability of at least one false discovery, but also on the probability of *any* discovery (true or false), whenever the treatment has a negative welfare effect on at least one outcome (i.e. $\theta \in \Theta_0^y$).

It is also possible to "invert" our research question and examine what researcher incentives and features of the publication process rationalize other popular criteria such as control of the FDR. Interestingly, it turns out that rationalizing FDR control in our model requires us to assume that the researcher is malevolent. In particular if

$$\beta_r^y(\theta) = \int \left[\sum_{g=1}^{G} \frac{1\left\{\theta_g < 0\right\} r_g(x)}{\sum_{g=1}^{G} r_g(x)} \cdot 1\left\{\sum_{g=1}^{G} r_g(x) > 0\right\}\right] dF_\theta(x) - C(G) \tag{20}$$

any decision rule that controls the FDR under the null hypothesis Θ_0^y at level $C(G)$ is maximin optimal. Equation (20) imposes that the researcher is malevolent in the sense that her utility is increasing in the number of false discoveries.

4 Additional forms of interactions between treatments

So far, we have analyzed settings where treatments interact only via the research cost function. Here we analyze two settings with successively more scope for interaction between hypotheses in each configuration. In the first we continue to assume no economic interactions between interventions but allow for interactions in the publication process through a threshold publication rule. In the second we allow for arbitrary economic interactions between interventions — for example, complementary interventions — as well as a threshold publication rule. We work from the model defined in Section 2 with perfectly informed experimenters, and in particular allow again for an arbitrary number of treatments. Appendix A.1 discusses settings with imperfectly informed experimenters.

4.1 Weak maximin rules

Maximin rules may be very conservative when we start considering additional forms of interactions between the treatments. Therefore, we introduce a weaker notion of maximin

optimality that considers the worst case over Θ_0 as in Definition 1 only (instead of Θ) and corresponds to the concept of *weak size control* in the MHT literature. Even under this weaker criterion we will often obtain conservative hypothesis testing protocols, which helps motivate attention to it.

Definition 5 (Weak maximin optimality)**.** We say that r^* is weakly maximin if and only if

$$r^* \in \arg\max_{r \in \mathcal{R}} \min_{\theta \in \Theta_0} v_r(\theta).$$

Following the same argument as in the proof of Proposition 1, we can show that r^* is weakly maximin if and only if[20]

$$\beta_{r^*}(\theta) \leq 0 \quad \text{for all } \theta \in \Theta_0. \tag{21}$$

Different from the notion of maximin optimality, Definition 5 considers the worst-case allocation over the set Θ_0 instead of Θ. It is a weaker notion of optimality since it requires size control only over the subset of parameters that lead to negative treatment effects for each possible recommendation function. That is, it imposes size control only under the weak null.[21] By definition, all maximin protocols are also weakly maximin, but the converse is not necessarily true. A weakly maximin protocol is also maximin only if welfare is weakly positive over $\Theta \setminus \Theta_0$.

4.2 Linear welfare and threshold-crossing publication rule

We now introduce interactions in the publication rule, replacing the linear rule in Assumption 2 with a threshold rule in which papers that find sufficiently many results can be published.

Assumption 5 (Threshold publication rule)**.** With a threshold publication rule, the experimenter's utility conditional on experimenting is (up-to-rescaling)

$$\beta_r(\theta) = \int \pi(r(x)) dF_\theta(x) \quad \pi(r(x)) = \begin{cases} \gamma - C(J) \text{ if } \sum_{j=1}^J r_j(x) \geq \kappa \\ -C(J) \text{ otherwise,} \end{cases} \tag{22}$$

for exogenous constants $\kappa \geq 0, \gamma > C(J)$.

[20] The formal argument goes as follows: any recommendation function r yields weakly negative welfare for $\theta \in \Theta_0$. Therefore the maximin recommendation function achieves zero utility over $\theta \in \Theta_0$, which holds if Equation (21) holds. This proves the "if" direction. If Equation (21) does not hold, then any r achieves negative utility, proving the "only if" direction.

[21] See Proschan and Brittain (2020) for related notions in the context of MHT.

The publication rule in Assumption 5 is analogous to the threshold crossing rule discussed in Section 3.2. The difference is that here $r(X)$ represents tests of the effects of different treatments on one outcome, whereas $r(X)$ represents tests of the effects of one treatment on multiple outcomes in Section 3.2. As before, we restrict the parameters so that there is at least one case in which it is profitable for the experimenter to conduct experimentation ($\gamma > C(J)$).

The threshold crossing publication rule leads to optimal hypothesis testing protocols which depend on the joint distribution of X in a complicated way and, thus, are difficult to interpret; see Appendix A.5 for a discussion. We will therefore restrict attention to the class of independent recommendation functions defined in Assumption 6 below.

Assumption 6 (Independent recommendations)**.** Consider a class of recommendation functions $\widetilde{\mathcal{R}} \subset \mathcal{R}$ with $r_j(X) \perp r_{j'}(X)$ with $j \neq j'$.

Assumption 6 states that tests for distinct treatments are statistically independent. This holds under the normality Assumption 4 when Σ is a diagonal matrix and $r_j(X)$ is a function of X_j only. Assumption 6 allows for separating the interactions arising from the threshold crossing publication rule from those occurring because of the statistical dependence between the components of X.[22]

The following proposition characterizes the ϵ-most powerful maximin recommendation functions under independence.

Proposition 6 (Partial adjustment with linear welfare and threshold publication rule)**.** *Let $J > 1$. Let Assumptions 1, 3, 4, 5, and 6 hold, and let $\Theta = [-1, 1]^J$. Then any $r^* \in \mathcal{R}$ is weakly maximin optimal and ϵ-most powerful if and only if r^* satisfies Equation (21) and*

$$\begin{aligned} &\lim_{\theta \downarrow 0} P\Big(r_j^*(X) = 1|\theta\Big) = P(r_j^*(X) = 1|\theta = 0) = p^* \quad \forall j \in \{1, \ldots, J\}, \\ &\text{where } p^* : \sum_{k \in \{\kappa, \cdots, J\}} \binom{J}{k} (p^*)^k = C(J)/\gamma, \end{aligned} \tag{23}$$

assuming such r^* exists.

Proof. See Appendix B.2.6. We note that the proof does not rely on normality of X (Assumption 4). We only require that X is continuously distributed with CDF F_θ, which admits a PDF $f_\theta(x)$ that is continuous in θ for all $x \in \mathcal{X}$. □

[22] Independence assumptions have been commonly used as a starting point for developing approaches to multiple testing (e.g., Benjamini and Liu, 1999; Finner and Roters, 2001), and provide an interesting benchmark for contrasting our results against existing procedures and recommendations.

Proposition 6 states that the optimal recommendation function involves separate size control and assigns to each false discovery the same probability p^*, which depends both on the number of hypotheses J and the threshold number of rejections κ needed for publication.

An immediate implication of Proposition 6 is that under Assumption 4, the threshold crossing rule of the form

$$r_j^*(X) = 1\Big\{X/\sqrt{\Sigma_{j,j}} \geq \Phi^{-1}(1-p^*)\Big\}, \quad \text{for all } j \in \{1,\ldots,J\},$$

is (weakly) maximin optimal. Here, the threshold depends on the size prescribed in the above proposition.

As in the case of the linear publication rule, a central implication of Proposition 6 is that the way the size p^* of hypothesis tests should vary with the number J of hypotheses tested depends on the structure of the research cost function $C(J)$. Whenever $C(J)/\gamma = \alpha$, which is equivalent to assuming a constant publication probability and constant costs in the number of discoveries, we can show that $p^* \asymp 1/J$ as $J \to \infty$. Therefore, fixed-cost research production functions ($C(J)/\gamma = \alpha$) again rationalize Bonferroni-style corrections; see Appendix A.5.

Figure 2 plots the optimal level of size control for different values of J. It shows a comparison between the optimal level of size control under a linear publication rule (Lemma 1) and threshold crossing publication rule (Proposition 6) with $\gamma = J$. We find that for any finite J the comparison of optimal test size under linear and threshold publication rules is ambiguous, depending on J and on the research cost function.

As Figure 2 illustrates, the optimal level of size control with a threshold crossing publication rule also depends on the location of the publication threshold κ. One can show that p^* is increasing in the threshold κ. Intuitively, as the threshold increases, it becomes harder for the experimenter to publish, and larger incentives are necessary to guarantee experimentation. As a result, for large-enough κ and fixed J, standard levels of size control such as 10% or 5% may be too stringent.

4.3 General welfare and threshold-crossing publication rule

We now introduce the possibility of economic interactions between the interventions being studied, in addition to interactions in the cost function $C(J)$ and the publication rule.

Assumption 7 (General welfare)**.** $u_j(\theta) = \theta_j$ for all j, with $\Theta = [-1,1]^{2^J-1}$.

Importantly, unlike Assumption 1, Assumption 7 allows for interactions in the welfare impact of multiple treatments. With $J = 2$, for example, complementarities can be modeled

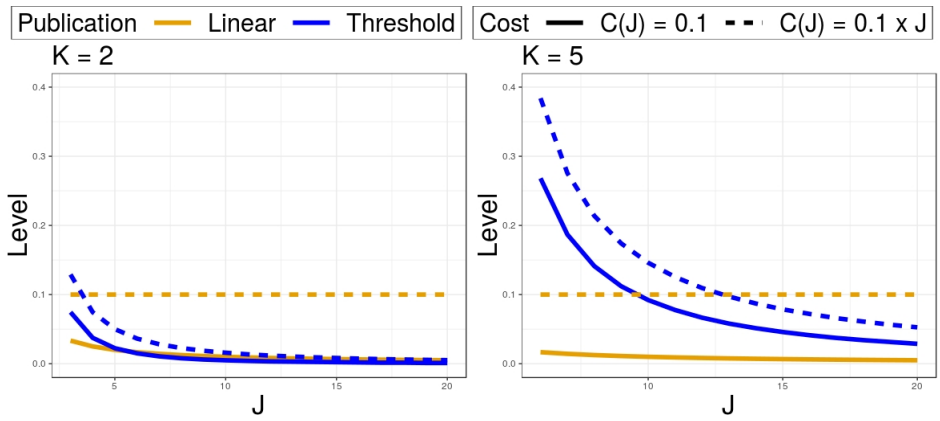


Figure 2: Linear welfare. Comparison of the optimal level of size control under linear and threshold publication rules. We set $C(J) \in \{0.1, 0.1 \times J\}$, where for the threshold crossing rule we fix $\gamma = J$. Different panels correspond to different values of κ for the threshold rule.

as $u_1(\theta) = \theta_1$, $u_2(\theta) = \theta_2$, and $u_3(\theta) = \theta_3 = \theta_1 + \theta_2 + \zeta$ for some $\zeta > 0$. As a result, we have $X \in \mathcal{X} \subseteq \mathbb{R}^{2^J-1}$ since we are interested in all possible combinations of treatments. We illustrate this more general setup in the context of our running example.

Example 3 (Fully saturated regression model)**.** Consider the fully saturated regression model

$$Y_i = D_{i,1}(1 - D_{i,2})\theta_1 + D_{i,2}(1 - D_{i,1})\theta_2 + D_{i,1}D_{i,2}\theta_3 + \varepsilon_i. \tag{24}$$

Unlike the "short" model (2), the "long" regression model (24) allows for interaction effects between the treatments, and θ_3 will differ from $\theta_1 + \theta_2$ in general. In this example, each entry of X corresponds to the OLS estimator of the effect of a *combination* of treatments, $X = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^\top$. □

To state the results we define some additional notation. Let

$$\widetilde{\delta}(r(X)) \in \{0,1\}^{2^J-1}, \quad \widetilde{\delta}_j(r(X)) = \begin{cases} \delta_j(r(X)) \text{ if } \sum_{j=1}^J r_j(X) \geq \kappa \\ 0 \text{ otherwise} \end{cases} \tag{25}$$

indicate the policy decisions taken *only* in the case that the experiment results in a paper that is publishable given the threshold publication rule.

The next assumption is a generalization of Assumption 3.

Assumption 8 (General tie-breaking assumption)**.** Assume that, conditional on X and for any $r \in \mathcal{R}$, the experimenter does not write the paper (or equivalently, reports $r(X) = (0, \ldots, 0)^\top$) if $\sum_{k=1}^K r_k(X) < \kappa$.

We can interpret Assumption 8 as imposing some infinitesimally small cost on the writing and submission process. If the experimenter knows that the paper will not be accepted after having seen X, she will not write up and submit the paper. We therefore consider the following modified version of the local alternative space.

Definition 6 (General ϵ-alternatives)**.** For $\epsilon > 0$, define the local alternative space as

$$\Theta_1(\epsilon) := \Big\{\theta : u_j(\theta) \geq \epsilon \text{ for some } j \in \mathcal{K}, u_j(\theta) \geq 0, \text{ for all } j\Big\},$$

where $\mathcal{K}$ denotes the set of indexes $k \in \{1, \ldots, 2^J - 1\}$ that corresponds to groups of κ or more hypotheses. □

Under Assumption 8, the editor only needs to consider groups of hypotheses for which it is profitable for the experimenter to conduct research. Therefore, we can write the editor and experimenter utility as

$$\beta_r(\theta) = \gamma \int \widetilde{\delta}(r(x))^\top \mathbf{1} dF_\theta(x) - C(J) \quad \text{and} \quad v_r(\theta) = \int \widetilde{\delta}(r(x))^\top u(\theta) dF_\theta(x). \tag{26}$$

The expressions for $\beta_r(\theta)$ and $v_r(\theta)$ in Equation (26) incorporate that whenever the number discoveries does not exceed the threshold κ, the experimenter does not submit the article such that the status-quo prevails.

The next proposition characterizes the optimal recommendation function for any given $\widetilde{\delta}_j$, which aggregates separate discoveries into a single recommendation. To state the proposition, we assume existence of an optimal solution. Such optimal solutions may not exist in general, and existence will depend on the distribution of X.

Proposition 7 (Equal size control on compound error rates)**.** *Let $J > 1$. Suppose that Assumptions 3, 4, 5, 7, and 8 hold. Suppose further that $\Theta = [-1, 1]^{2^J - 1}$. Let $\mathcal{K}$ be as in Definition 3. Then any $r^* \in \mathcal{R}$ is weakly maximin optimal and ϵ-most powerful if and only if r^* satisfies Equation (21) and*

$$P\Big(\widetilde{\delta}_j(r^*(X)) = 1 | \theta = 0\Big) = \frac{C(J)}{\gamma |\mathcal{K}|} \quad \forall j \in \mathcal{K}, \tag{27}$$

assuming such r^* exists.

Proof. See Appendix B.2.7. We note that the proof does not rely on normality of X (Assumption 4). We only require that X is continuously distributed with CDF F_θ, which admits a PDF $f_\theta(x)$ that is continuous in θ for all $x \in \mathcal{X}$. □

Proposition 7 shows that separate size control over each *group of discoveries* is maximin optimal and ϵ-most powerful whenever the effect of each group of discoveries has equal weight on the experimenter's utility. It rationalizes a specific form of FWER control.

Corollary 1 (Rationalization of the weak FWER)**.** *Let the conditions in Proposition 7 hold. Then any MHT procedure that satisfies Equation (27) controls the weak FWER at level $C(J)/\gamma$, namely*

$$P\left(\widetilde{\delta}_j(r^*(X)) = 1 \text{ for at least one } j|\theta = 0\right) = C(J)/\gamma \quad \forall j \in \mathcal{K}.$$

Proof. See Appendix B.3.1. □

In other words, Proposition 7 rationalizes (weak) FWER control between *groups* of hypotheses sufficient for publication. Importantly, FWER is not applied to each separate discovery $r_j(X)$, but instead to each group. When a single result is sufficient for publication ($\kappa = 1$), however, this implies control of the probability of a single false rejection, i.e. of the standard notion of weak FWER control. This follows from the fact that rejecting any group of hypotheses implies rejecting its constituent members individually (for $\kappa = 1$, $\max_j \delta_j(r(X)) = 1$ if and only if $\max_j r_j(X) = 1$).

Next we provide an example of a hypothesis testing protocol that satisfies the conditions in Proposition 7.

Example 4 (Recommendation function with $|\mathcal{K}|$ and independence)**.** Consider a vector $X \in \mathbb{R}^{|\mathcal{K}|}$, with each entry corresponding to a statistic X_j corresponding to a certain *group* of treatments with number of discoveries exceeding κ.[23] Assume that $X \sim \mathcal{N}(\theta, I)$, then the recommendation function[24]

$$\widetilde{\delta}_j(r(X)) = 1\left\{X_j > \max_{j' \neq j} X_{j'} \text{ and } X_j > t\right\} \tag{28}$$

is maximin and ϵ-most powerful if t is chosen such that $P\left(\max_j X_j > t|\theta = 0\right) = C(J)/\gamma$.[25] The recommendation function (28) is maximin optimal since

$$P\left(\max_j X_j > t|\theta\right) = P\left(\max_j (X_j + \theta_j) > t|\theta = 0\right) \leq P\left(\max_j X_j > t|\theta = 0\right)$$

for any $\theta \in \Theta_0 = \{\theta \in \Theta : \theta < 0\}$. It is ϵ-most powerful because

$$\begin{aligned} &P\left(1\left\{X_j > \max_{j' \neq j} X_{j'} \text{ and } X_j > t\right\}|\theta = 0\right) \\ &= P\left(\max_{j'} X_{j'} > t|\theta = 0, \max_{j'} X_{j'} \leq X_j\right) P\left(\max_{j'} X_{j'} \leq X_j|\theta = 0\right). \end{aligned}$$

□

[23] Consider Example 3 where $J = 2$: if $\kappa = 1$, then $|\mathcal{K}| = 2^2 - 1 = 3$; if instead, $\kappa = 2$, then $|\mathcal{K}| = 1$.

[24] This recommendation function bears some resemblance with step-down procedures in Lehmann and Romano (2005b, Chapter 9), where the maximum is considered a statistic of interest.

[25] We note that a simple threshold crossing rule violates the constraint that $\sum_j \delta_j(r(X)) \leq 1$.

Example 4 provides an example of a maximin and ϵ-most powerful recommendation function. The independence between the entries of X is important here; without it the existence of an optimal recommendation function is not guaranteed. Example 4 also illustrates the complexity of optimal recommendation functions in the presence of potential interactions between interventions.

5 Discussion & implications for practice

A theme throughout the analysis has been that MHT adjustments may be warranted when hypothesis tests interact in important ways. While it is intuitive that these include interactions between the economic effects of the interventions being tested such as those in Section 4.3, we have seen in Section 2 that even without these interactions adjustments may be appropriate when there are interactions in the cost of doing research. We close by discussing the implications of these latter results for applied empirical work in practice and the implications of our results for problems with multiple outcomes. We focus on broad and intuitive themes that seem likely to hold in a variety of plausible models of the research publication process, rather than on the specific functional forms that emerge as solutions to our specific model.

First, **multiple hypothesis testing protocols should reflect the structure of research costs.** This is the central idea in Section 2.5 (Lemma 1 and Proposition 4). To illustrate the practical implications more concretely, it is useful to decompose (without loss of generality) the total research costs $C(J)$ into fixed and variable components c_f and $c_v(J)$:

$$C(J) = c_f + c_v(J). \tag{29}$$

Using this notation, the optimal level of size control in Lemma 1 is $C(J)/J = (c_f + c_v(J))/J$.

Consider first a case in which the research production function exhibits no economies of scale with respect to the number of hypotheses being tested. Specifically, suppose there are no fixed costs and that variable costs scale linearly in J, i.e. $c_f = 0$ and $c_v(J) = \alpha J$. Then it is a direct corollary of Proposition 4 that r^* is maximin optimal and ϵ-most powerful if r^* satisfies Equation (12) with threshold t such that

$$P(r_j^*(X) = 1 | \theta = 0) = \alpha \quad \forall j \in \{1, \ldots, J\}. \tag{30}$$

In other words, standard inference without adjustment for MHT is optimal in this case. While it is true that the researcher obtains a higher expected reward from taking on projects that test more hypotheses, the appropriate correction for this is already "built in" to the costs of conducting research, so that no further correction is required.

Now consider the case in which the research production function exhibits strong economies of scale. Specifically, suppose there are only fixed costs and that the marginal cost of testing an additional hypothesis is zero, i.e. $c_v(J) = 0$ and $c_f = \alpha$. Then r^* is maximin optimal and ϵ-most powerful if r^* satisfies Equation (12) with threshold t such that

$$P(r_j^*(X) = 1|\theta = 0) = \frac{\alpha}{J}, \quad \forall j \in \{1, \ldots, J\}. \tag{31}$$

In this case the optimal inferential procedure is to control average size, as for example via a Bonferroni correction. This correction is necessary to appropriately align the incentives of the researcher with those of the editor, as without it she would have a disproportionate incentive to conduct projects with a large number of hypotheses.

In practice, the structure of the research cost function may often lie somewhere in-between these two extremes. Testing an additional intervention usually increases the costs of running an experiment, for example, but may well do so less than proportionately. Our framework would then motivate forms of multiple testing correction that are weaker than Bonferroni adjustments. There could even be *dis-economies* of scale in some research production processes, in which case our framework yields the interesting implication that hypothesis testing procedures should *reward* rather than penalize researchers for testing multiple hypotheses. The key, over-arching point is that the framework pins down the optimal correction as a function of quantities that are potentially observable, using data on the costs of various research projects and the numbers of hypotheses they test.

Second, **multiple *outcomes* may or may not warrant multiple testing adjustments depending on the breadth of their intended audience.** Multiplicity of outcomes is of particular practical interest, as it is very common. Program evaluation papers, for example, typically examine many more outcomes than they do interventions or subpopulations.

Editors in our framework should consider requiring MHT adjustments to tests of multiple outcomes when the research is expected to inform a broad, heterogeneous audience of policy-makers who are likely to value different outcomes differently. In this case, and if the editor wishes to maximize the welfare of the worst-off policy-maker, MHT adjustment across multiple outcomes can be attractive. But editors in our framework should not require MHT adjustments simply because a paper reports multiple outcomes. If these outcomes all speak to a single decision — as for example in a program evaluation that informs whether or not to scale up the program — then separate inference on the individual outcomes is not relevant to that decision. The policy-maker's decision will instead be based on the overall welfare effect of the program, which is an aggregate of the individual effects.

The practical implication is that editors should seek to obtain information on this aggre-

gation. One ambitious approach would be to change the nature of publication process itself, creating interactive electronic formats that allow individual readers to input the values they attach to various outcomes themselves and obtain inference on the implied indices automatically. A more incremental reform would be to simply require that researchers elicit such valuations from a relevant policy-maker and test the resulting index. This approach would be similar in spirit to the increasingly popular practice of testing for effects on researcher-defined index variables that summarize a group of related outcomes. An important difference, however, is that the relevant weights used to construct the index would not be based on purely statistical properties of the outcomes (as for example in principal components analysis, or the inverse (co)variance weighting approach of Anderson (2008)). In some cases this approach can lead to results that make little economic sense—down-weighting both children's health and education, for example, because the two outcomes are highly correlated. Our framework calls instead for indexing using weights that appropriately reflect each outcome's economic significance.

References

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

Andrews, I. and J. M. Shapiro (2020). A model of scientific communication. NBER Working Paper.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and W. Liu (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference 82*(1-2), 163–170.

Chassang, S., G. Padro I Miquel, and E. Snowberg (2012). Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review 102*(4), 1279–1309.

Di Tillio, A., M. Ottaviani, and P. N. Sørensen (2017). Persuasion bias in science: Can economics help? *The Economic Journal 127*(605), F266–F304.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science 23*(1), 1–22.

Efron, B. et al. (2008). Simultaneous inference: when should hypothesis testing problems be combined? *Annals of Applied Statistics 2*(1), 197–223.

Elliott, G., U. K. Müller, and M. W. Watson (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica 83*(2), 771–811.

Fink, G., M. McConnell, and S. Vollmer (2014). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness 6*(1), 44–57.

Finner, H. and M. Roters (2001). On the false discovery rate and expected type i errors. *Biometrical Journal 43*(8), 985–1005.

Frankel, A. and M. Kasy (2018). Which findings should be published? Working Paper.

Gaivoronski, A. (1986). Linearization methods for optimization of functionals which depend on probability measures. In *Stochastic Programming 84 Part II*, pp. 157–181. Springer.

Henry, E. and M. Ottaviani (2019). Research and the approval process: the organization of persuasion. *American Economic Review 109*(3), 911–55.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2020). Chapter 4 - Asymptotic analysis of statistical decision rules in econometrics. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin (Eds.), *Handbook of Econometrics, Volume 7A*, Volume 7 of *Handbook of Econometrics*, pp. 283–354. Elsevier.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics 6*, 65–70.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

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

Lee, S. and A. M. Shaikh (2014). Multiple testing and heterogeneous treatment effects: re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics 29*(4), 612–626.

Lehmann, E. L. and J. P. Romano (2005a). Generalizations of the familywise error rate. *The Annals of Statistics 33*(3), 1138–1154.

Lehmann, E. L. and J. P. Romano (2005b). *Testing statistical hypotheses*. Springer Science & Business Media.

Lehmann, E. L., J. P. Romano, and J. P. Shaffer (2005). On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics 33*(3), 1084 – 1108.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*(4), 773–793.

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

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

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

Proschan, M. A. and E. H. Brittain (2020). A primer on strong vs weak control of familywise error rate. *Statistics in medicine 39*(9), 1407–1413.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California.

Romano, J. P., A. Shaikh, and M. Wolf (2011). Consonance and the closure method in multiple testing. *The International Journal of Biostatistics 7*(1).

Romano, J. P., A. M. Shaikh, and M. Wolf (2010). Hypothesis testing in econometrics. *Annu. Rev. Econ. 2*(1), 75–104.

Romano, J. P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters 113*, 38–40.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

Storey, J. D. et al. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics 31*(6), 2013–2035.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Tetenov, A. (2016). An economic theory of statistical testing. Working Paper.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wald, A. (1950). *Statistical decision functions*. Wiley.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.

Williams, C. (2021). Preregistration and incentives. SSRN 3796813.

Yoder, N. (2019). Designing incentives for heterogeneous researchers. SSRN 3154143.

Appendix to "(When) should you adjust inferences for multiple hypothesis testing?"

A Extensions **1**
A.1 Imperfectly informed experimenters 1
A.2 One-sided and two-sided hypothesis testing 4
A.3 Σ-robust recommendation functions 4
A.4 Multiple outcomes, single policy-maker, adversarial weights 5
A.5 Additional details Section 4.2 7

B Proofs **8**
B.1 Lemmas 9
B.2 Propositions 11
B.3 Corollaries 25

A Extensions

A.1 Imperfectly informed experimenters

In Section 2, we assume that the researcher has perfect information and knows θ. Here we show that our main results continue to hold in settings with imperfect information.[26]

Let $\pi \in \Pi$ denote the prior of the experimenter about the value of θ, where Π is the class of *all* distributions supported on Θ. Throughout this section, we assume that Π is unrestricted.[27] The prior π represents knowledge about θ that is available to both the experimenter and the policy-maker, but not to the editor. This assumption reflects the fact that the researcher and the policy-maker typically know more about the specific interventions than the editor.[28]

We assume that the vector of statistics X is drawn from a normal distribution conditional on θ, where θ itself is drawn from the prior π. Formally, the data generating process is as

[26] In the single-hypothesis testing case, Tetenov (2016) gives results under imperfect information. However, these results rely on the Neyman-Pearson lemma, which is not applicable in our setting with multiple testing.

[27] This assumption is made for simplicity. For our theoretical results, we only need that the class of priors Π contains at least one element that is supported on the null space Θ_0, which holds by construction if Π is unrestricted.

[28] If the prior π was known to the editor, she would act as a Bayesian decision-maker.

follows:

$$X \mid \theta \sim \mathcal{N}(\theta, \Sigma), \quad \theta \sim \pi, \quad \pi \in \Pi,$$

where Σ is positive definite.

In this setup, the ex-ante experimenter utility and social welfare are both functions of the prior π and defined as

$$\begin{aligned} \widetilde{\beta}_r(\pi) &= \int \beta_r(\theta) d\pi(\theta), \\ \widetilde{v}_r(\pi) &= \begin{cases} \int\int u(\theta)^\top \delta(r(x)) dF_\theta(x) d\pi(\theta) & \text{if } \widetilde{\beta}_r(\pi) > 0 \\ \max\left\{ \int\int u(\theta)^\top \delta(r(x)) dF_\theta(x) d\pi(\theta), 0 \right\} & \text{if } \widetilde{\beta}_r(\pi) = 0 \\ 0 & \text{if } \widetilde{\beta}_r(\pi) < 0. \end{cases} \end{aligned} \tag{32}$$

The experimenter acts as a Bayesian decision maker and maximizes her decision of whether to experiment. The decision is based on whether her expected utility, $\widetilde{\beta}_r(\pi)$, is positive. The policy-maker incurs a loss (or revenue) depending on the experimenter's decision, which also depends on the function r chosen by the editor.

Under imperfect information, we define maximin optimal recommendation functions as follows.

Definition 7 (Π-maximin optimal)**.** We say that $\tilde{r}^*$ is Π-maximin optimal if and only if

$$\tilde{r}^* \in \arg\max_{r \in \mathcal{R}} \inf_{\pi \in \Pi} \widetilde{v}_r(\pi).$$

In Definition 7, we define maximin rules with respect to the prior π, which is known to the experimenter and policy-maker. The definition generalizes the previous notion of maximin optimality with respect to the parameters θ (as opposed to π). When Π contains only point mass distributions, Definition 7 is equivalent to maximin optimality considered in Section 2.3.

The following lemma provides a characterization of maximin optimality.

Lemma 2 (Conditions for maximin optimality)**.** *Let $\Theta_0 \neq \emptyset$ as in Definition 1. The recommendation function $\widetilde{r}^*$ is Π-maximin optimal if and only if $\inf_{\pi \in \Pi} \widetilde{v}_{\widetilde{r}^*}(\pi) \geq 0$.*

Proof. See Appendix B.1.2. □

Lemma 2 states that maximin optimality is equivalent to the worst-case policy-maker objective function being non-negative. Based on this result, the next proposition shows that one-sided t-tests with appropriately chosen critical values are maximin optimal and admissible under imperfect information.

Proposition 8 (Maximin optimality and admissibility)**.** *Let $J > 1$. Let Assumption 1, 2, 3, and 4 hold, and let $\Theta = [-1, 1]^J$,*[29] *$C(J) > 0$. Then the recommendation function*

$$\widetilde{r}_j^*(X) = 1\Big\{X_j/\sqrt{\Sigma_{j,j}} \geq \Phi^{-1}\left(1 - C(J)/J\right)\Big\}, \quad \forall j \in \{1, \cdots, J\}$$

is Π-maximin optimal. In addition, r^* is also admissible with respect to any $\pi \in \Pi$.

Proof. See Appendix B.2.8. The proof of the first result (maximin optimality) uses the duality properties of the linear program. The second result instead is a consequence of Proposition 4. □

Proposition 8 shows that the conclusions in Section 2.5 on the maximin optimality of t-tests remain valid under imperfect information. Proposition 8 further states that r^* is admissible. While maximin optimality connects to size control in hypothesis testing, admissibility captures a notion of power. It implies that we cannot find any other decision rule that is more powerful than r^* for all distributions.

Remark 4 (ϵ-most powerful rules under imperfect information)**.** Unlike the result under perfect information in Proposition 4, we do not characterize ϵ-most powerful rules in Proposition 8. This is because with imperfect information, there are several different notions of ϵ-most powerful tests. For instance, we may consider alternatives that assign ϵ probability to a certain parameter value in the alternative space or alternatives that assign probability one to the parameter taking value ϵ. The first notion implicitly imposes a certain prior on a specific value of the parameter, which may be hard to justify in practice. The second notion coincides with the one discussed in the main text and is satisfied by the decision rule in Proposition 8. □

Our results for imperfectly informed experimenters can be extended to the settings with general welfare functions and threshold crossing publication rules as studied in Sections 4.2 and 4.3. In these sections, maximin optimality is defined with respect to Θ_0 only. We can impose an equivalent condition under imperfect information with respect to

$$\Pi^{\text{weak}} = \Big\{\pi : \int_{\theta \in \Theta_0} \pi(\theta) d\theta = 1\Big\}.$$

The set Π^{weak} denotes the set of priors that impose mass one on the null space Θ_0. The notion of weak maximin optimality and the characterization of weakly maximin decisions is the same as the one in Equation (21). To see why, note that any decision rule that satisfies $\beta_r(\theta) \leq 0, \forall \theta \in \Theta_0$, also satisfies $\int \beta_r(\theta) d\pi(\theta) \leq 0$ for all $\pi \in \Pi^{\text{weak}}$.

[29] We note that it suffices that Θ is compact for the result to hold.

A.2 One-sided and two-sided hypothesis testing

The structure of the game we study naturally justifies one-sided hypothesis testing. Researchers test whether or not a proposed treatment strictly improves upon baseline. Treatments that lead to negative values of welfare are effectively excluded from consideration since the editor discourages experimentation in this case. This structure is a direct consequence of the assumption that the policy-maker always implements the recommended treatment. That said, our framework justifies two-sided hypothesis testing under appropriate changes to the structure of the game. We outline these modifications here but leave the formal analysis of the resulting game for future research.

Consider a policy-maker who randomly selects a treatment to implement using a uniform distribution when no recommendation is made. For each treatment, the researcher can make three recommendations: (i) implement the treatment; (ii) not implement the treatment; (iii) "do not know", which corresponds to the baseline, i.e., implement the treatment with some prior probability. Denote the recommendation for treatment j as $\tilde{r}_j(X) \in \{-1, 0, 1\}$, with $\tilde{r}_j(X) = -1$ corresponding to not implementing the treatment, $\tilde{r}_j(X) = 0$ corresponding to no recommendation, and $\tilde{r}_j(X) = 1$ corresponding to implementing the treatment. Consider a linear model without complementarities between treatments. The welfare generated by the recommendation $\tilde{r}_j(X)$ is

$$\tilde{v}_{\tilde{r}}(\theta) = \sum_{j=1}^{J} \int \left(1\Big\{\tilde{r}_j(x) = 1\Big\}\theta_{j,1} + 1\Big\{\tilde{r}_j(x) = -1\Big\}\theta_{j,-1} \right) dF_\theta(x).$$

Here $\theta_{j,1}, \theta_{j,-1}$ can take arbitrary values. The coefficients $\theta_{j,1}, \theta_{j,-1}$ capture the benefit net of the *opportunity* cost of implementing the treatment and not implementing the treatment respectively, relative to the random baseline. For the case where the policy-maker implements no treatment as baseline, then $\theta_{j,-1} = 0$, recovering our model formulation. By assuming that the baseline intervention is a random intervention, the model justifies *two-sided* hypothesis testing: not implementing a treatment has a benefit (or cost) relative to random implementation.

A.3 Σ-robust recommendation functions

In the main text, we assume that the experimental design and sample size (and thus the covariance matrix of X, Σ) are known to the editor. In this setting, the editor chooses the recommendation function r to maximize worst-case welfare given Σ. Here we analyze a variant of our model in which the editor chooses r when Σ unknown and adversarially chosen

by nature. We refer to recommendation functions that are maximin optimal in this setting as Σ-robust.

Definition 8 (Σ-robust)**.** We say that r^* is Σ-robust if

$$r^* \in \arg\max_{r \in \mathcal{R}} \min_{\theta, \Sigma_{j,j} > \kappa \forall j} v_r(\theta), \quad \text{for some } \kappa > 0.$$

Definition 8 states that the rule r^* is Σ-robust if it is maximin optimal not only with respect to θ but also with respect to Σ. The definition imposes a lower bound κ on the diagonal elements of the covariance matrix, ensuring that the signal-to-noise ratio $\theta_j/\sqrt{\Sigma_{j,j}}$ is uniformly bounded. This lower bound can be interpreted as a lower bound on the sample size necessary to publish a paper.

Σ-robust recommendation functions are attractive in settings where journals impose consistent universal editorial policies and norms that do not depend on the specifics of the experimental design.

The next proposition shows that threshold crossing rules are Σ-robust.

Proposition 9. *Let $J > 1$. Let Assumption 1, 2, 3, and 4 hold, and let $\Theta = [-1, 1]^J$. Then the recommendation function*

$$\widetilde{r}_j^*(X) = 1\Big\{X_j/\sqrt{\Sigma_{j,j}} \geq \Phi^{-1}\left(1 - C(J)/J\right)\Big\}, \quad \forall j \in \{1, \ldots, J\}$$

is Σ-robust.

Proof. The proof mimics the proof of Proposition 4 (see Appendix B.2.4), since the optimization problem only depends on $\theta_j/\sqrt{\Sigma_{j,j}}$, and is omitted. □

Proposition 9 shows that threshold crossing rules and standard t-tests are also optimal in settings where the editor seeks recommendation functions that are optimal irrespectively of the particular experimental design.

A.4 Multiple outcomes, single policy-maker, adversarial weights

Here we analyze a setup similar to the one in Section 3.1, but with adversarial weights. Namely, suppose that the welfare weights w^* are unknown to the editor (but not the policy-maker) and — in the spirit of the maximin criteria we studied above — chosen adversarially by nature. This assumption reflects information asymmetry between the researcher and policy-maker on the one hand and the editor on the other. As in the main text, we assume that the policy-maker passively implements the recommendation.

The adversarial welfare criterion is

$$v_r(\theta; w^{adv}), \quad w^{adv} \in \arg\min_{w\in\Delta} \sum_{g=1}^{G} w_g u_g(\theta), \tag{33}$$

where Δ denotes the G-simplex. The worst-case welfare conditional on the researcher experimenting can be written as

$$\begin{aligned}\int \min_{w\in\Delta} r(x) \sum_{g=1}^{G} w_g u_g(\theta) dF_X(\theta) &= \int r(x) dF_X(\theta) \sum_{g=1}^{G} w_g^{adv} u_g(\theta) \\ &= \int r(x) dF_X(\theta) \min\Big\{u_1(\theta), \cdots, u_G(\theta)\Big\}.\end{aligned}$$

The experimenter's utility only depends on publication prospects and costs, and on the welfare and the weights. The null space takes the following form:

$$\Theta_0^y := \{\theta : u^g(\theta) < 0, \text{ for some } g \in \{1, \ldots, G\}\} \tag{34}$$

where we use the superscript y to make explicit that this is the null space in the case of multiple outcomes. The definition of the null space is similar to Section 3.2.[30]

The next proposition characterizes the set of maximin recommendation functions with adversarial welfare weights.

Proposition 10 (Adversarial weights)**.** *Let Assumption 3 hold and suppose that $\Theta_0^y \neq \emptyset$. A recommendation function r^* is maximin-optimal with adversarial weights i.e.*

$$r^* \in \arg\max_{r\in\mathcal{R}} \min_{\theta\in\Theta} v_r(\theta; w^{adv})$$

if and only if

$$\beta_{r^*}(\theta) \leq 0 \quad \forall \theta \in \Theta_0^y, \tag{35}$$

where Θ_0^y is defined in Equation (34).

Proof. See Appendix B.2.9. □

Note that the structure of the result is similar to that in Proposition 1 in Tetenov (2016); the key difference is that null set within which the researcher is deterred from experimenting is potentially much *larger*.

The next corollary provides an example of a maximin recommendation function in this context. As in Section 2.2 we write $C := C(1)$. Recall that here $r(X)$ is a scalar.[31]

[30] It is interesting to note that, as a result, the problem in this subsection and the one in Section 3.2 lead to similar conclusions.

[31] This is different from Section 3.2 where $r(X)$ is a $G \times 1$ vector.

Corollary 2 (Threshold crossing recommendation function with maximin protocols over outcomes)**.** *Suppose that*

$$X_g \sim \mathcal{N}\Big(\theta_g, 1\Big), \quad \theta_1, \ldots, \theta_G \in [-M, M], \quad X_g \perp X_{g' \neq g}.$$

Define $X := \min\{X_1, \ldots, X_G\}$*. Consider a linear publication rule as in Lemma 1, and a protocol* $r(X;t) = 1\{X \geq t\}$*. Then* $r^*(\cdot) := r(\cdot; t^*)$ *is maximin optimal, where* t^* *satisfies* $P(X_g \geq t^* | \theta_g = 0) = C$*. In addition, for some* M *large enough, any protocol* $r(X;t)$ *with* $P(X_g \geq t | \theta = 0) > C$ *is not maximin.*

Proof. See Appendix B.3.2. □

Corollary 2 characterizes a particular class of maximin protocols. Interestingly, the corollary shows that under independence, by choosing $X = \min\{X_1, \ldots, X_G\}$, a maximin threshold recommendation function imposes a very stringent size control of the form

$$P(r(X) = 1 | \theta = 0) = C^G, \tag{36}$$

Note that $C < 1$ if the researcher experiments, so that this implies the size of the test shrinks with respect to the number of outcomes at an exponential rate. This adjustment arises due to the adversarial nature of the game with unknown weights and the independence assumption.

Corollary 2 also states that any threshold crossing recommendation function that violates the above size control is not maximin optimal.[32] That said, there may exist an alternative class of hypothesis testing protocols that lead to higher power as G increases; we leave this for future study.

A.5 Additional details Section 4.2

Motivating Assumption 6. To motivate the independence Assumption 6, consider the leading case where $X \sim \mathcal{N}(\theta, \Sigma)$. One can show[33] that any weakly maximin and ϵ-most powerful recommendation function satisfies

$$P(r_j^*(X) = 1 | \theta = 0) \geq p^*, \quad p^* = \min\{p_1^*, \ldots, p_J^*\}, \tag{37}$$

[32] The idea of the proof is that by taking $X = \min_g X_g$, we violate maximin optimality whenever each parameter except for one is large and positive, and for one g' only $\theta_{g'}$ is close to zero but negative. In such a case the probability of a discovery, which under independence is obtained by taking the product of each $P(X_g > t|\theta)$ exceeds the overall cost C and thus induces experimentation.

[33] The proof follows similarly to the proof of Lemma 1 where the worst case alternative puts mass ϵ on the discovery with the smallest probability. As a result, the editor wants to maximize the minimum probability across all discoveries.

where $(p_1^*, \ldots, p_J^*)$ are the solutions to the following optimization problem

$$
\begin{aligned}
&(p_1^*, \ldots, p_J^*) \in \arg \max_{p \in [0,1]^J} \min_{j \in \{1,\ldots,J\}} p_j \\
&\text{such that } \sum_{k \in \mathcal{K}} P(\delta_k(r(X)) = 1|\theta) \leq C(J)/\gamma \quad \forall \theta \in \Theta_0 \\
&\text{and } P(r_j(X) = 1|\theta = 0) = p_j \quad \forall j \in \{1, \cdots, J\}.
\end{aligned}
$$

The above expression shows that admissible recommendation functions impose that the probability of discovery of each separate treatment exceeds a certain (uniform) threshold p^*. The threshold depends on the joint distribution of the entries of X, which rules out separate size control if there is dependence between the entries of X.

Asymptotic relationship between p^* and J in Proposition 6. We characterize asymptotics in two cases of interest: one where $C(J)/\gamma$ is constant, and one where $C(J)/\gamma = 1/J$. The first case is essentially equivalent to assuming a constant publication probability and constant costs in the number of discoveries. The second case corresponds to assuming that the experimenter utility is increasing in J (e.g., the quality of the publication increases in the number of tests performed).

Corollary 3 (Asymptotic approximation)**.** *Assume that κ is fixed. Let p^* be as defined in Proposition 6. Suppose that $C(J)/\gamma = \alpha$ for a constant $1 > \alpha > 0$ independent of J. Then $p^* \asymp 1/J$ as $J \to \infty$. Suppose instead that $C(J)/\gamma = \alpha/J$. Then $p^* \asymp 1/J^{(\kappa+1)/\kappa}$ as $J \to \infty$.*

Proof. See Appendix B.3.3. □

Corollary 3 shows that fixed-cost research production functions ($C(J)/\gamma = \alpha$) again rationalize Bonferroni-style corrections. Here for a threshold publication rule this holds asymptotically, as opposed to the linear publication rule case in which the result was exact. Interestingly, when $C(J)/\gamma = \alpha/J$, size control is of order $1/J^2$ for $\kappa = 1$ and approximately $1/J$ for $\kappa \gg 1$.

B Proofs

In our proofs we will sometimes write C instead of $C(J)$ and $P(r_j(X)|\theta)$ instead of $P(r_j(X) = 1|\theta)$ to lighten up the notation whenever it does not cause any confusion. For notational convenience, without loss of generality, whenever possible, we standardize the threshold utility with respect to γ, taking $\gamma = 1$. To distinguish maximin from weakly maximin, we refer to maximin as strongly maximin.

B.1 Lemmas

Here we collect the proofs of all lemmas.

B.1.1 Proof of Lemma 1

We prove the statement in the following steps. First observe that we can write the experimenter's payoffs *proportional to*

$$\sum_{j=1}^{J} P(r_j(X)|\theta) - C. \tag{38}$$

The proof proceeds in the following steps. We first find a lower bound on the worst-case power of r^*. We then argue that any weakly maximin recommendation function that does not satisfy the conditions of r^* has a lower power using an upper bound on any maximin $r \neq r^*$.

Step 1: Maximin optimality. This directly follows from Proposition 1.

Step 2: ϵ-more powerful calculations We claim that

$$\liminf_{\epsilon \downarrow 0} \frac{1}{\epsilon} \inf_{\theta \in \Theta_1(\epsilon)} v_{r^*}(\theta) \geq \frac{C}{J}.$$

We now show why. Define $\theta(\epsilon) \in \Theta_1(\epsilon)$ the vector of parameter under the local alternative. Observe that the utility under the local alternative reads as follows

$$(A) = \inf_{\theta(\epsilon)} \sum_{j=1}^{J} \theta(\epsilon) P(r_j^*(X)|\theta = \theta(\epsilon)), \text{ such that } \theta_j(\epsilon) \geq \epsilon \text{ for some } j, \quad \theta_j(\epsilon) \geq 0 \quad \forall j.$$

We then write

$$\begin{aligned}(A) &\geq \inf_{w \in [0,1]^J : \sum_j w_j \geq \epsilon, \theta(\epsilon) \in \Theta_1(\epsilon)} \sum_{j=1}^{J} w_j P(r_j^*(X)|\theta = \theta(\epsilon)) \\ &\geq \inf_{w \in [0,1]^J : \sum_j w_j \geq \epsilon, \theta' \in [0,1]^J : \sum_j \theta_j \geq \epsilon} \sum_{j} w_j P(r_j^*(X)|\theta = \theta') := g(\epsilon).\end{aligned}$$

Define $\mathcal{W}(\epsilon_1, \epsilon_2) = \left\{(w, \theta) \in [0,1]^{2J} : \sum_j w_j \geq \epsilon_1, \sum_j \theta_j \geq \epsilon_2\right\}$. Observe that we can write

$$\frac{1}{\epsilon} g(\epsilon) = \inf_{(w,\theta') \in \mathcal{W}(1,\epsilon)} \sum_{j=1}^{J} w_j P(r_j^*(X)|\theta = \theta') = \inf_{(w,\theta') \in \mathcal{W}(1,1)} \sum_{j=1}^{J} w_j P(r_j^*(X)|\theta = \epsilon\theta').$$

Observe that $\mathcal{W}(1,1)$ is a compact space. In addition $P(r_j^*(X)|\theta = \epsilon\theta')$ is continuous in ϵ for any $\theta' \in \Theta$ by Assumption 4. As a result, $g(\epsilon)/\epsilon$ is a continuous function in ϵ. Therefore, we obtain that

$$\lim_{\epsilon\to 0}\frac{g(\epsilon)}{\epsilon} = \inf_{(w,\theta')\in\mathcal{W}(1,1)}\sum_{j=1}^{J} w_j P(r_j^*(X)|\theta = \theta' \times 0) = \inf_{(w,\theta')\in\mathcal{W}(1,1)}\sum_{j=1}^{J} w_j \frac{C}{J} = \frac{C}{J}.$$

This completes the proof of our claim.

Step 3: Alternative set of maximin protocols. We now claim that any maximin protocol which is not r^* must satisfy for *some* $j \in \{1, \cdots, J\}$,

$$P\Big(r_j(X) = 1|\theta_k = 0, \quad \forall k \in \{1, \cdots, J\}\Big) < \frac{C}{J}. \tag{39}$$

The claim holds for the following reasons. Consider a maximin recommendation function r' such that for all j Equation (49) does not hold. Then if r' is maximin optimal and satisfies Equation (39) with equality for all j, then there must be an r^* defined as in the proposition statement equal to r', which leads to a contradiction. Therefore it must be that r' is such that for some j Equation (39) is satisfies with reversed *strict* inequality and for all j is satisfied with reversed weak inequality. To observe why, by the above argument

$$\sum_j P\Big(r_j'(X) = 1|\theta_k = 0, \quad \forall k \in \{1, \cdots, J\}\Big) > C.$$

As a result, take $\theta = (-t, -t, \cdots, -t) \in \Theta_0$ for some small t. Then by Assumption 4 (namely, by continuity of F_θ), we have for t small enough

$$\sum_j P\Big(r_j'(X) = 1|\theta_k = -t, \quad \forall k \in \{1, \cdots, J\}\Big) > C.$$

As a result, for t small enough, the policy r' contradicts Proposition 1.

Step 4: Power comparison. Observe now that for the class of alternative treatments, we have

$$\inf_{\theta(\epsilon)}\sum_{j=1}^{J}\theta(\epsilon)P(r_j(X)|\theta = \theta(\epsilon)) \leq \epsilon P(r_j(X) = 1|\theta_j = \epsilon, \theta_{-j} = 0),$$

since the allocation $(\theta_j = \epsilon, \theta_{-j} = 0) \in \Theta_1(\epsilon)$. Using Assumption 4 we have

$$\lim_{\epsilon\to 0} P(r_j(X) = 1|\theta_j = \epsilon, \theta_{-j} = 0) = P(r_j(X) = 1|\theta = 0) < \frac{C}{J}.$$

This completes the proof of the if statement.

Step 5: "Only if" statement The only if statement follows from the fact that if r^* does not satisfy the condition in the proposition, then we can find a different function r' which leads to larger power than r^* following the same argument after Equation (39). As a result, in this case r^* violates the condition of local optimality.

B.1.2 Proof of Lemma 2

First, recall that by construction $v_r(\theta) \leq 0$, for all $\theta \in \Theta_0$ and all r. As a result, for all r, $\int v_r(\theta) s\pi(\theta) \leq 0$, for all $\pi \in \Pi$, such that $\int_{\theta \in \Theta_0} \pi(\theta) d\theta = 1$. Since there exists at least one π which puts probability one on Θ_0, it follows that $\inf_{\pi \in \Pi} \widetilde{v}_r(\pi) \leq 0$ for all r. Therefore, r is maximin optimal if $\inf_{\pi \in \Pi} \widetilde{v}_r(\pi) \geq 0$.

Note now that by choosing $r^t(X) = (0, \cdots, 0)$ almost surely, we are guaranteed that $\inf_{\pi \in \Pi} \widetilde{v}_r(\pi) = 0$. Therefore, r^* is maximin optimal only if $\inf_{\pi \in \Delta} \widetilde{v}_r(\pi) \geq 0$, since otherwise dominated by r^t.

B.2 Propositions

Here we collect the proofs of all propositions.

B.2.1 Proof of Proposition 1

Let $\Theta_0^s = \Theta \setminus \left\{\theta : u_j(\theta) \geq 0 \forall j\right\}$. We start with a general observation. Define $\theta^*(r) = \min_{\theta \in \Theta} v_r(\theta)$ the worst-case θ as a function of the recommendation function r. First, observe that since $\Theta_0 \neq \emptyset$, we have that $v_r(\theta) \leq 0$ for any $r \in \mathcal{R}, \theta \in \Theta_0$. Therefore, it must be that any recommendation function is maximin if and only if $v_{r^*}(\theta^*(r^*)) = 0$, since (i)if $v_{r^*}(\theta^*(r^*)) < 0$, then the editor can choose $\tilde{r}(X) = (0, 0, \ldots, 0)$ and obtain $v_{\tilde{r}(X)}(\theta) = 0, \forall \theta \in \Theta$; (ii) if instead $v_{r^*}(\theta^*(r^*)) > 0$, then we reach a contradiction since we can find a $\theta \in \Theta_0$ which leads to non-positive utility. This shows that maximin rules can equivalently be characterized by $v_{r^*}(\theta^*(r^*)) = 0$.

Based on this observation, to prove the "if" direction, we only need to show that the worst-case utility under r^* equals zero, that is $v_{r^*}(\theta^*(r^*)) = 0$. Under Assumption 3, the editor's payoffs equal exactly zero for all $\theta \in \Theta_0$, as long as Equation (9) is satisfied. To complete the first direction of the claim, we are left to show that for any $\theta \in \Theta \setminus \Theta_0$ the editor's payoff is always non-negative. This is true since if $\theta \notin \Theta_0^s$ then the utility is trivially positive, since $u(\theta)$ has all entries weakly positive. If instead $\theta \in \Theta_0^s$, then either (i) $\theta \in \Theta_0$ or (ii) $\theta \in \Theta_0^s \setminus \Theta_0$. (i) was discussed before. Therefore consider (ii). Observe that since $v_{r^*}(\tilde{\theta}) \geq 0$ for all $\tilde{\theta} \in \Theta_0^s \setminus \Theta_0$ by assumption, it must be that $v_{r^*}(\theta^*(r^*)) = 0$. As a result if Equation (9) holds, r^* is maximin.

We now discuss the "only if" direction. Consider the case where $\beta_r(\theta) > 0$ for some $\theta \in \Theta_0$. Then we have $v_r(\theta^*(r)) \leq v_r(\theta) < 0$ for some $\theta \in \Theta_0$. Suppose instead that $v_{r^*}(\tilde{\theta}) < 0$, for some $\tilde{\theta} \in \Theta \setminus \Theta_0$. Then similarly $v_r(\theta^*(r)) \leq v_r(\tilde{\theta}) < 0$, completing the proof.

B.2.2 Proof of Proposition 2

We can write the experimenter's payoffs *proportional to*

$$\sum_{j=1}^{J} P(r_j(X)|\theta) - C, \tag{40}$$

where we suppress the dependence of C with J for notational convenience. We take the parameter space

$$\Theta = \Big\{\theta \in [-1,1]^J \text{ such that } \text{sign}(\theta_1) = \text{sign}(\theta_2) = \cdots = \text{sign}(\theta_{J-1})\Big\}$$

and the costs $C/(J-1) \leq 1$, which is possible since $J > 1$. corresponding to the positive and negative quadrants for all parameters except θ_J which can have an arbitrary sign irrespective of the sign of the other parameters. To prove the statement we show that there exists a maximin protocol that strictly dominates all others over an arbitrary set $\Theta' \subseteq \Theta$, and a different maximin recommendation function that it strictly dominates all others over some arbitrary set $\Theta'' \neq \Theta', \quad \Theta'' \subseteq \Theta$. We choose $\Theta' = (0, \cdots, 0, t)$ for a small t and $\Theta'' = (t, \cdots, t, 0)$ for a small t. Details are discussed in the following paragraphs. Since we can choose any distribution F_θ, we choose

$$X \sim \mathcal{N}(\theta, I).$$

Observe that we can define

$$\Theta_0 \subseteq \tilde{\Theta}_0 = \Big\{\theta : \theta_j \leq 0 \text{ for all } j\Big\},$$

where $\tilde{\Theta}_0$ also contains those elements that lead the editor's utility to be *weakly* negative. We break the proof into several steps.

Step 1: construction of the function class. Define a class of recommendation functions of the form

$$\mathcal{R}^1 = \Big\{r \in \mathcal{R} : C \geq P\Big(r_J^1(X)|\theta = 0\Big) > \frac{C}{J} \text{ and } r \text{ is maximin } \Big\}.$$

We claim that $\mathcal{R}^1 \neq \emptyset$. To prove this claim it suffices to find a function $r \in \mathcal{R}^1$. An example is

$$\begin{aligned} P\Big(r_j(X)\Big|\forall\theta \in \Theta\Big) &= 0 \quad \forall j \leq J-1, \\ P\Big(r_J(X)|\forall\theta_J = 0, \forall\theta_{j<J} \leq 0\Big) &= \min\{C, 1\}, \end{aligned} \tag{41}$$

and $P(r_J(X)|\theta_J, \theta_{-J})$ is constant in θ_{-J} and decreasing in θ_J, which exist since $\frac{C}{J} < 1$. r is maximin optimal since the experimenter's utility is weakly non-negative for any $\theta \in \Theta$.[34] This follows by (i) monotonicity of $P(r_J(X) = 1|\theta_J, \theta_{-J})$ in θ_J and (ii) the fact that $P(r_J(X) = 1|\theta_J, \theta_{-J})$ is constant in θ_{-J}. A simple example satisfying such conditions is a threshold crossing recommendation function of the form

$$r_j(X) = 1\{X_j > t_j\}, t_{j<J} = \infty, t_J = \Phi^{-1}(1 - \min C, 1)$$

where $\Phi(\cdot)$ denotes the normal CDF. In addition, we observe that

$$\sup_{r \in \mathcal{R}^1} P\Big(r_J(X)|\theta_J = 0, \theta_{j<J} = 0\Big) \geq \min\{C, 1\}$$

as a result of the above example. Define $\tau = \min\{C, 1\}$ for the rest of the proof.

Step 2: comparisons with maximin protocols. We now claim that for $\theta = (0, 0, \cdots, 0, t)$, for t approaching zero, there exists a maximin recommendation function $r^1 \in \mathcal{R}^1$ which leads to strictly larger editor's utility than any *maximin* decisions $r^2 \in \mathcal{R} \setminus \mathcal{R}^1$. To show our claim, it suffices to compare r^1, to any recommendation function $r^2 \notin \mathcal{R}^1$, such that

$$P\Big(r_J^2(X)|\theta = 0\Big) \leq \frac{C}{J}. \tag{42}$$

To see why, observe that whenever the above probability is between $(\frac{C}{J}, \tau]$, we contradict the statement that $r^2 \notin \mathcal{R}^1$. When instead

$$P\Big(r_J^2(X)|\theta = 0\Big) > \tau$$

the recommendation function r^2 is *not* maximin optimal, since this implies that $C \leq 1$, which in turn implies that by Assumption 4, the experimenter would conduct experimentation under $(\theta_J = -t, \theta_{j<J} = -t)$, for some small positive t, leading to strictly negative editor's utility.

Step 3: Comparisons of editor's utility. Observe now that for $\theta = (0, 0, \cdots, 0, t)$, the editor's utility reads as follows

$$\frac{1}{t} v_{r^1}(0, 0, \cdots, 0, t) = P\Big(r_J^1(X_1, X_2, \cdots, X_J)\Big|\theta_{-J} = 0, \theta_J = t\Big) = f_{r_J^1}(t).$$

[34] See the proof of Proposition 1 for an explanation why a weakly non negative utility for all $\theta \in \Theta$ implies maximin optimality.

Notice that $f_r(t)$ is a continuous function in t due to Assumption 4. By comparing the utilities under r^1 and r^2, and taking the difference we have

$$\sup_{r^1\in\mathcal{R}^1} v_{r^1}(\theta_J = t, \theta_{j<J} = 0) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} v_{r^2}(\theta_J = t, \theta_{j<J} = 0) =$$
$$t\Big[\sup_{r^1\in\mathcal{R}^1} f_{r^1_J}(t) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} f_{r^2_J}(t)\Big] =$$
$$t\Big[\sup_{r^1\in\mathcal{R}^1} f_{r^1_J}(0) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} f_{r^2_J}(0) + \varepsilon_{r^1_J}(t) - \varepsilon_{r^2_J}(t)\Big]$$

where $\lim_{t\to 0}\varepsilon_{r^1_J}(t) = \lim_{t\to 0}\varepsilon_{r^2_J}(t) = 0$, by Assumption 4. As a result, we can take some small $t > 0$, such that

$$\varepsilon_{r^1_J}(t) - \varepsilon_{r^2_J}(t) < \sup_{r^1\in\mathcal{R}^1} f_{r^1_J}(t) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} f_{r^2_J}(t),$$

since $\sup_{r^1\in\mathcal{R}^1} f_{r^1_J}(t) > \frac{C}{J}$, and therefore $\sup_{r^1\in\mathcal{R}^1} f_{r^1_J}(t) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} f_{r^2_J}(t) > 0$. For this case, we have

$$\sup_{r^1\in\mathcal{R}^1} v_{r^1}(\theta_J = t, \theta_{j<J} = 0) - \sup_{r^2: P(r^2_J(X)|\theta=0)\leq \frac{C}{J}} v_{r^2}(\theta_J = t, \theta_{j<J} = 0) > 0.$$

Therefore, for some t small enough, r^1 leads to strictly larger utility than any maximin recommendation function $r^2 \notin \mathcal{R}^1$.

Step 4: the recommendation function r^1 is not Pareto dominant We are left to show that there exists a function $r^3 \notin \mathcal{R}^1$ which is maximin optimal and that leads to strictly larger utility than any $r^1 \in \mathcal{R}^1$ for some different combinations of θ. We choose $\theta = (t, t, \cdots, 0)$. We construct a set of decisions

$$\mathcal{R}^2 = \Big\{r \notin \mathcal{R}^1 \text{ and } r \text{ is maximin optimal}\Big\}.$$

Observe that $\mathcal{R}^1 \cap \mathcal{R}^2 = \emptyset$ by definition.

Step 5: welfare computation for $\mathcal{R}^2$. We claim that $\mathcal{R}^2$ is non-empty. An example is the treatment that assigns $P(r_J(X) = 1|\theta = 0) = 0$,

$$P(r_{j<J}(X) = 1|\theta_j = 0, \quad \forall\theta_{-j}) = \frac{C}{(J-1)},$$

which $P(r_j(X) = 1|\theta_j, \theta_{-j})$ decreasing in θ_j and does not depend on θ_{-j}, for $j \neq J$. A threshold crossing recommendation function satisfies this condition, since $\frac{C}{(J-1)} \leq 1$. Such

a recommendation function is maximin by the assumption on Θ and the monotonicity of $P(r_j(X)=1|\theta_j)$ in θ_j. Consider the alternative $\check{\theta}=(t,\cdots,t,0)$. Observe now that we have

$$\sup_{s^2\in\mathcal{R}^2} v_{s^2}(\check{\theta})-\sup_{r^1\in\mathcal{R}^1} v_{r^1}(\check{\theta})=t\times\Big[\sum_{j<J}P(s_j^2(X)=1|\theta=\check{\theta})-\sum_{j<J}P(r_j^1(X)=1|\theta=\check{\theta})\Big].$$

We write

$$\begin{aligned}&\Big[\sum_{j<J}P(s_j^2(X)=1|\theta=\check{\theta})-\sum_{j<J}P(r_j^1(X)=1|\theta=\check{\theta})\Big]\\&=\Big[\sum_{j<J}P(s_j^2(X)=1|\theta=0)-\sum_{j<J}P(r_j^1(X)=1|\theta=0)\Big]+\epsilon(t)\end{aligned}$$

where $\epsilon(t)\to 0$ as $t\to 0$ by continuity (Assumption 4).

Step 6: Upper bound on r^1. We claim that

$$\sum_{j<J}P(r_j^1(X)=1|\theta=0)<C.$$

We prove the claim by contradiction. Suppose that the above equation does not hold. Then it must be that

$$\sum_{j<J}P(r_j^1(X)=1|\theta=0)+P(r_J^1(X)=1|\theta=0)>C+\frac{C}{J},\tag{43}$$

which contradicts maximin optimality of r^1. Using continuity, we obtain that for t small enough any $r^1\in\mathcal{R}^1$ is dominated by r^2. The proof is complete.

B.2.3 Proof of Proposition 3

We structure the proof as follows: we first prove the first part of the claim. We show that any decision which is strongly maximin and ϵ-more powerful than any other strongly maximin protocol is admissible. Second, we show that such decision is also admissible once compared to recommendation functions which are not maximin. Finally, we prove the second part of the claim (the only if statement).

We assume a fully interacted utility of the form $u_j(\theta)=\theta_j$ throughout our proof. The proof in the linear case follows similarly.

Step 1: ϵ-more powerful implies admissibility among strongly maximin protocols. Observe first that we can take $\theta_j=\epsilon,\theta_{i\neq j}=0$ which belongs to $\Theta_1(\epsilon)$ since by assumption $\theta_j\in[-b,b]$ for some positive b. Observe that by the definition of ϵ-more powerful

$$0\leq\inf_{\theta\in\Theta_1(\epsilon)}v_r(\theta)\leq P(r_j(X)=1|\theta_j=\epsilon,\theta_{-j}=0)\epsilon\leq\epsilon\tag{44}$$

for some j, where the right-hand side follows by definition of $\Theta_1(\epsilon)$. As a result we have that $0 \leq \inf_{\theta \in \Theta_1(\epsilon)} v_r(\theta)/\epsilon \leq 1$ and so the lim-inf and lim-sup are uniformly bounded. We now have that

$$(I) = \liminf_{\epsilon \downarrow 0} \left\{ \frac{1}{\epsilon} \inf_{\theta \in \Theta_1(\epsilon)} v_r(\theta) - \frac{1}{\epsilon} \inf_{\theta' \in \Theta_1(\epsilon)} v_{r'}(\theta') \right\}$$

is finite since by Equation (44) the above expression for any $\epsilon > 0$ is bounded from below by zero and from above by one. Now observe that by definition of lim-inf, there exists a *subsequence* $\epsilon_n \downarrow 0$, such that $\frac{1}{\epsilon_n} \inf_{\theta \in \Theta_1(\epsilon_n)} v_r(\theta) - \frac{1}{\epsilon_n} \inf_{\theta' \in \Theta_1(\epsilon_n)} v_{r'}(\theta')$ converges to (I). Take some finite n over the subsequence but large enough such that

$$\frac{1}{\epsilon_n} \inf_{\theta \in \Theta_1(\epsilon_n)} v_r(\theta) - \frac{1}{\epsilon_n} \inf_{\theta' \in \Theta_1(\epsilon_n)} v_{r'}(\theta') > 0.$$

Define $\theta_{\epsilon_n} = \inf_{\theta' \in \Theta_1(\epsilon_n)} v_{r'}(\theta')$. Observe that

$$\frac{1}{\epsilon_n} v_r(\theta_{\epsilon_n}) - \frac{1}{\epsilon_n} v_{r'}(\theta_{\epsilon_n}) \geq \frac{1}{\epsilon_n} \inf_{\theta \in \Theta_1(\epsilon_n)} v_r(\theta) - \frac{1}{\epsilon_n} \inf_{\theta' \in \Theta_1(\epsilon_n)} v_{r'}(\theta') > 0, \tag{45}$$

since[35]

$$\inf_{\theta \in \Theta_1(\epsilon_n)} v_r(\theta) \leq v_r(\theta_{\epsilon_n}).$$

Now observe that Equation (45) implies that r strictly dominates r' at some θ_{ϵ_n}. We now want to show that r is admissible within the class of maximin protocols. Since r is more ϵ-powerful of all maximin protocols, we can apply the same reasoning to any other r'' for some (different) θ'_{ϵ_n}. Clearly, r is not dominated by any other maximin protocol recommendation function by the argument above.

Step 2: Comparison to non-maximin recommendation function. Next we show that the strongly maximin ϵ-more powerful r is also admissible within the larger class of strongly maximin and non-strongly maximin protocols. To show this we use a contradiction argument. Suppose that there exists a recommendation function $\tilde{r}$ which is *not* maximin Then $\tilde{r}$ is dominated by r either over the parameter $\theta \in \Theta_0$ or over $\Theta \setminus \Theta_0$ since r is maximin. Therefore r is admissible.

B.2.4 Proof of Proposition 4

Observe that

$$P(r_j(X) = 1 | \theta = 0) = \frac{C}{J}, \quad \forall j \in \{1, \cdots, J\}.$$

[35] The inequality below follows from the fact that the constraint set $\Theta_1(\epsilon)$ does not depend on r.

As a result, we only need to prove maximin optimality. Observe first that $r_j(X)$ is monotonically increasing in θ_j and constant in θ_{-j}. In addition, the null is defined as

$$\Theta_0 \subseteq \Big\{\theta : \theta_j \leq 0, \quad \forall j\Big\}.$$

By monotonicity, the first condition in Proposition 1 is satisfied. We show that also the second condition holds.

To show this it suffices to show that the worst-case objective function is weakly positive. With an abuse of notation we define θ_j the coefficient rescaled by $\sqrt{\Sigma_{j,j}}$ (which are finite by the assumption of Σ being positive-definite). We write the nature's adversarial game with a threshold crossing recommendation function as follows

$$\min_{\theta\in\Theta} \sum_{j=1}^{J}(1-\Phi(t-\theta_j))\theta_j, \qquad \text{s.t.} \sum_{j=1}^{J}(1-\Phi(t-\theta_j)) \geq J(1-\Phi(t)). \tag{46}$$

Observe that since $C(J) > 0$, t must be finite. We can write Θ as a compact space $[-M, M]^J$ for some finite M, by assumption:

$$\min_{\theta\in[-M,M]^J} \sum_{j=1}^{J}(1-\Phi(t-\theta_j))\theta_j, \qquad \text{s.t.} \sum_{j=1}^{J}(1-\Phi(t-\theta_j)) \geq J(1-\Phi(t)). \tag{47}$$

for some arbitrary large M whose choice is discussed below. Observe that Equation (47) is weakly smaller than Equation (46) hence a lower bound on Equation (47) suffices to prove the claim.

Claim. We claim that the editor utility can only be negative if the minimizer θ^* is such that for some j, $\theta_j^* < 0$ and for some other $j' \neq j$ $\theta_{j'}^* > 0$. That there must exist some negative θ_j^* trivially follows from the objective function. That there must be a positive $\theta_{j'}^*$ follows directly from the constraint function: if such condition is not met and $\theta_j^* < 0$ for all j, then it follows that the constraint is violated.

Focus on interior solution for M large enough. We now argue that at least one component of the minimizer must satisfy $-\infty < \theta_j^* < 0$ for the resulting objective function to be strictly negative. To show this it suffices to observe that $z(1-\Phi(t-z)) \to 0$ as $z \to -\infty$. Therefore if for all $\theta_j^* < 0$ these are unbounded, then the objective function is trivially zero proving the claim. Second, following this same argument we also observe that it suffices to focus our analysis on solutions $\theta^* \in \Theta \subseteq [-M, M]^J$ for some arbitrary large but finite M. To see why observe that if *at least one* θ_j^* is unbounded its contribution to the objective function is zero, while it decreases the experimenter's utility $\sum_{j=1}^{J}(1-\Phi(t-\theta_j^*))$,

hence having a weakly positive effect on the objective function. For the $\theta_j^* > 0$ these must instead be finite since otherwise the objective function is strictly positive (hence, the claim trivially holds). Hence there must exist a minimizer θ^* which is in the interior of $[-M, M]^J$ for some arbitrary large and finite M.

Constraint qualification. We now show that the KKT conditions are necessary for the optimality of $\theta^* \in \Theta \subset [-M, M]^J$. To show this we use the LICQ. In particular observe that the derivative of the constraint function is

$$-\sum_{j=1}^{J} \phi(t - \theta_j) \neq 0$$

for t being finite, for any point θ such that at least one θ_j is finite (in absolute value). If such condition is not met, then the objective function is (weakly) non negative as discussed in the previous paragraph.

Lagrangian. We now study *necessary* conditions for the optimal solution of the problem in Equation (46). Consider the Lagrangian function

$$\sum_{j=1}^{J}(1 - \Phi(t - \theta_j))\theta_j + \lambda\Big[J(1 - \Phi(t)) - \sum_{j=1}^{J}(1 - \Phi(t - \theta_j))\Big] + \mu_{1,j}[\theta - M] + \mu_{2,j}[-\theta - M].$$

Now observe that by the argument in the second paragraph, and complementary slackness we can focus on the cases where $\mu_{1,j} = \mu_{2,j} = 0$ for all j (i.e. θ^* is an interior of $[-M, M]^J$ for some finite M large enough). Taking first order conditions of the Lagrangian we obtain

$$\phi(t - \theta_j)\theta_j + (1 - \Phi(t - \theta_j)) = \lambda\phi(t - \theta_j) \Rightarrow \lambda^* = \frac{1}{\phi(t - \theta_j)}\Big[\phi(t - \theta_j)\theta_j + (1 - \Phi(t - \theta_j))\Big].$$

Contradiction argument. We conclude this proof using a contradiction argument using the claim we established at the beginning of the proof. Suppose that the objective function is strictly negative. The there must exists a j such that $\theta_j^* < 0$ and $j' \neq j$ such that $\theta_{j'}^* > 0$. In addition, observe that using the equation for the optimal λ, we can write

$$0 > \theta_j^* = \lambda^* - \frac{(1 - \Phi(t - \theta_j^*))}{\phi(t - \theta_j^*)} \quad 0 < \theta_{j'}^* = \lambda^* - \frac{(1 - \Phi(t - \theta_{j'}^*))}{\phi(t - \theta_{j'}^*)}.$$

Using the fact that t is finite, it follows that

$$\frac{1 - \Phi(t - \theta_{j'}^*)}{\phi(t - \theta_{j'}^*)} < \lambda^* < \frac{1 - \Phi(t - \theta_j^*)}{\phi(t - \theta_j^*)}.$$

Observe now that the expression implies

$$\frac{1-\Phi(z)}{\phi(z)} < \frac{1-\Phi(z')}{\phi(z')} \text{ for some } z < z'.$$

However, by standard properties of the normal CDF $(1-\Phi(z))/\phi(z)$ is monotonically *decreasing* in z hence leading to a contradiction.

B.2.5 Proof of Proposition 5

The proof mimics the proof of Proposition 1. Observe that since $\Theta_0^y \neq \emptyset$, there exist a $\theta \in \Theta_0^y$ such that $v_r(\theta) \leq 0$ for any $r \in \mathcal{R}$. Then the maximin utility equals zero, since otherwise the editor can set $r(X) = 0$ and achieve zero utility. Observe that the editor payoffs is negative for all $\theta \in \Theta_0^y$. Under Assumption 3, the editor's payoffs equal exactly zero for all $\theta \in \Theta_0$, as long as Equation (35) is satisfied. In addition, the editor payoff's is always non-negative for $\theta \in \Theta \setminus \Theta_0^y$ and any r, for any $\theta \setminus \Theta_0^y$ each $u^g(\theta)$ is weakly positive. As a result if Equation (35) holds, r^* is maximin. Consider now the case where $\beta_r(\theta) > 0$ for some $\theta \in \Theta_0^y$. Then, we can find a $\theta \in \Theta_0^y$ which leads to negative utility, for any decision $r(\cdot)$. Assumption 3 is invoked for the equality in Equation (35) to be a weak inequality.

B.2.6 Proof of Proposition 6

For the sake of brevity we refer to maximin protocols as weakly maximin. The proof of the theorem follows similarly to Lemma 1. First, observe that the experimenter's utility is linear in $\delta_k(r(X))$ for every $k \in \mathcal{K}$, since $\sum_{k\in\mathcal{K}} \delta_k(r(X)) \leq 1$ and $\delta_k(r(X)) \in \{0,1\}$. Namely, the experimenter's utility reads as follows

$$\sum_{k\in\mathcal{K}} P(\delta_k(r(X)) = 1|\theta) - C(J),$$

where $\mathcal{K}$ denotes the set of discoveries exceeding κ recommendations, where we rescaled $\gamma = 1$. We first prove the first part of the statement.

Step 1: Maximin optimality. Maximin optimality directly follows from Equation (21).

Step 2: ϵ-more powerful calculations. We *claim* that

$$\liminf_{\epsilon\downarrow 0} \frac{1}{\epsilon} \inf_{\theta\in\Theta_1(\epsilon)} v_{r^*}(\theta) \geq p^*.$$

The claim holds following the same argument as in the proof of Lemma 1.

Step 3: Alternative set of maximin protocols We now claim that any maximin protocol which is not r^* must satisfy for *some* $j \in \{1, \cdots, J\}$,

$$P\Big(r'_j(X) = 1 | \theta_k = 0, \quad \forall k\Big) < p^*. \tag{48}$$

The claim follows directly from the incentive-compatibility constraint, following the same argument as in Lemma 1, with $P(\delta_j(r(X))|\theta) = p^{*k}$ for any group j having k many treatments (all treatments must be selected). There are $\binom{J}{k}$ many groups having k treatments. As a result, we can find a worst-case allocation for $r'_j(X)$ which leads to a utility bounded from above by ϵp^*. The rest of the proof follows similarly to Lemma 1 and Proposition 7 below.

B.2.7 Proof of Proposition 7

For the sake of brevity we refer to maximin protocols as weakly maximin. Observe that since the experimenter never submits discoveries with less than κ treatments (Assumption 8), we only focus on discoveries having a positive effect on the experimenter's utility. Therefore we refer to $\sum_k$ as $\sum_{k \in \mathcal{K}}$. Finally, observe that the experimenter's utility is linear in $\delta_k(r(X))$ for every $k \in \mathcal{K}$, since $\sum_{k \in \mathcal{K}} \delta_k(r(X)) \leq 1$ and $\delta_k(r(X)) \in \{0, 1\}$. As a result, we can prove the statement in the following steps following the proof of Lemma 1 with minor modifications. We first prove the first part of the statement.

Step 1: Maximin optimality. This directly follows from Equation (21).

Step 2: ϵ-more powerful calculations. We *claim* that

$$\liminf_{\epsilon \downarrow 0} \frac{1}{\epsilon} \inf_{\theta \in \Theta_1(\epsilon)} v_{r^*}(\theta) \geq \frac{C}{|\mathcal{K}|}.$$

We now show why. Define $\theta(\epsilon) \in \Theta_1(\epsilon)$ the vector of parameter under the local alternative. Observe that the utility under the local alternative reads as follows

$$(A) = \inf_{\theta(\epsilon)} \sum_{k \in \mathcal{K}} \theta(\epsilon) P(\delta_k(r^*(X)) = 1 | \theta = \theta(\epsilon)), \text{ such that } \theta_k(\epsilon) \geq \epsilon \text{ for some } k \in \mathcal{K}, \quad \theta_k(\epsilon) \geq 0 \quad \forall k \in \mathcal{K}.$$

We then write

$$\begin{aligned}(A) &\geq \inf_{w \in [0,1]^{|\mathcal{K}|} : \sum_k w_k \geq \epsilon, \theta(\epsilon) \in \Theta_1(\epsilon)} \sum_{k \in \mathcal{K}} w_k P(\delta_k(r^*(X)) = 1 | \theta = \theta(\epsilon)) \\ &\geq \inf_{w \in [0,1]^{|\mathcal{K}|} : \sum_k w_k \geq \epsilon, \theta \in [0,1]^{|\mathcal{K}|} : \sum_k \theta_k \geq \epsilon} \sum_k w_k P(\delta_k(r^*(X)) | \theta) := g(\epsilon).\end{aligned}$$

Define $\mathcal{W}(\epsilon_1, \epsilon_2) = \Big\{(w, \theta) \in [0,1]^{2\times|\mathcal{K}|} : \sum_j w_k \geq \epsilon_1, \sum_k \theta_k \geq \epsilon_2\Big\}$. Observe that we can write

$$\frac{1}{\epsilon} g(\epsilon) = \inf_{(w,\theta')\in\mathcal{W}(1,1)} \sum_{k\in\mathcal{K}} w_k P\Big(\delta_k(r^*(X))|\theta = \theta'\epsilon\Big).$$

Observe that $\mathcal{W}(1,1)$ is a compact space. In addition $P(\delta_k(r^*(X)) = 1|\theta)$ is continuous in θ by Assumption 4. As a result, $g(\epsilon)/\epsilon$ is a continuous function in ϵ. Therefore, we obtain that

$$\lim_{\epsilon\to 0} \frac{g(\epsilon)}{\epsilon} = \inf_{(w,\theta)\in\mathcal{W}(1,1)} \sum_{k\in\mathcal{K}} w_k P(\delta_k(r^*(X)) = 1|\theta \times 0).$$

Observe now that by construction of the recommendation function, the above expression equals to

$$\inf_{(w,\theta)\in\mathcal{W}(1,1)} \sum_{k\in\mathcal{K}} w_k P(\delta_k(r^*(X)) = 1|\theta \times 0) = \inf_{(w,\theta)\in\mathcal{W}(1,0)} \sum_{k\in\mathcal{K}} w_k \frac{C}{|\mathcal{K}|} = \frac{C}{|\mathcal{K}|}.$$

This completes the proof of our claim.

Step 3: Alternative set of maximin protocols. We now claim that any maximin protocol which is not r^* must satisfy for *some* $j \in \mathcal{K}$,

$$P\Big(\delta_j(r(X)) = 1|\theta_k = 0, \quad \forall k\Big) < \frac{C}{|\mathcal{K}|}. \tag{49}$$

The claim holds for the following reasons. Consider a maximin protocol r' such that for all k Equation (49) does not hold. Then if r' is maximin optimal and satisfies Equation (49) with equality for all k, then there must be an r^* defined as in the proposition statement equal to r. Therefore it must be that r' is such that for some k Equation (49) is satisfied with reversed *strict* inequality and for all k is satisfied with reversed weak inequality. To observe why, by the above argument

$$\sum_j P\Big(\delta_j(r'(X)) = 1|\theta_k = 0, \quad \forall k\Big) > C.$$

As a result, take $\theta = (-t, -t, \cdots, -t) \in \Theta_0$ for some small t. Then by Assumption 4 (namely, by continuity of F_θ), we have for t small enough

$$\sum_j P\Big(\delta_j(r'(X)) = 1|\theta_k = -t, \quad \forall k\Big) > C.$$

As a result, for t small enough, the protocol r' contradicts Proposition 1.

Step 4: Power comparison. Observe now that for the class of alternative treatments, we have

$$\inf_{\theta(\epsilon)} \sum_{k \in \mathcal{K}} \theta(\epsilon) P(\delta_k(r(X)) | \theta = \theta(\epsilon)) \leq \epsilon P(\delta_k(r(X)) = 1 | \theta_k = \epsilon, \theta_{-k} = 0),$$

since the allocation $(\theta_k = \epsilon, \theta_{-k} = 0) \in \Theta_1(\epsilon)$. Using Assumption 4 we have

$$\lim_{\epsilon \to 0} P(\delta_k(r(X)) = 1 | \theta_k = \epsilon, \theta_{-k} = 0) = P(\delta_k(r(X)) = 1 | \theta = 0) < \frac{C}{|\mathcal{K}|}.$$

This completes the "if" part of the statement. The "only if" part follows from the same argument used after Equation (49).

B.2.8 Proof of Proposition 8

We first show maximin optimality. To show maximin optimality, it suffices to show that the worst-case objective function is weakly positive. With an abuse of notation, we define θ_j the coefficient rescaled by $\sqrt{\Sigma_{j,j}}$.

Preliminaries for maximin optimality. We write the nature's adversarial game with a threshold crossing recommendation function as follows

$$\min_{\pi \in \Pi} \int \sum_{j=1}^{J} (1 - \Phi(t - \theta_j)) \theta_j d\pi(\theta), \quad \text{s.t.} \int \sum_{j=1}^{J} (1 - \Phi(t - \theta_j)) d\pi(\theta) \geq J(1 - \Phi(t)). \tag{50}$$

Observe that since $C(J) > 0$, t must be finite. It will be convenient to consider the equivalent optimization program

$$\max_{\pi \in \Pi} - \int \sum_{j=1}^{J} (1 - \Phi(t - \theta_j)) \theta_j d\pi(\theta), \quad \text{s.t.} \int \sum_{j=1}^{J} (1 - \Phi(t - \theta_j)) d\pi(\theta) \geq J(1 - \Phi(t)). \tag{51}$$

Here, we inverted the sign of the objective and consequently inverted the maximum with the minimum. We will show that Equation (51) is bounded from *above* by zero, which suffices to show maximin optimality.

Finite dimensional optimization program. The maximization over $\pi \in \Pi$ can be equivalently rewritten as a minimization over $\pi_1(\cdot), \cdots, \pi_J(\cdot), \theta_j \sim \pi_j$, i.e., with respect to marginal distributions. This follows directly by additivity. By Theorem 1, result 3 in Gaivoronski (1986), we can write the optimization problem in Equation (50) as an optimization

over some finitely many $n \times J$ discrete points $\left(\theta_1^i, \cdots, \theta_J^i\right)_{i=1}^n$, each point θ_j^i having marginal probability $p_{i,j}$.[36] Hence, we write

$$(50) = \max_{\pi \in \Pi_n} \sum_{i=1}^n \int \sum_{j=1}^J (\Phi(t - \theta_j^i) - 1)\theta_j p_{i,j}, \quad \text{s.t.} \sum_{i=1}^n \sum_{j=1}^J (\Phi(t) - \Phi(t - \theta_j^i))p_{i,j} \geq 0, \quad (52)$$

where

$$\Pi_n = \left\{ (p_{i,j})_{i=1,j=1}^{n,J}, p_{i,j} \geq 0, \sum_{i=1}^n p_{i,j} = 1 \text{ for all } j \in \{1, \cdots, J\} \right\},$$

for some $\left(\theta_1^i, \cdots, \theta_J^i\right)_{i=1}^n$.

Dual formulation. The optimization in Equation (52) is a linear program with linear constraints. Therefore, the dual is directly defined as follows.

$$\begin{aligned} &\min_{y_1, \cdots, y_J, y_{J+1}} \sum_{j=1}^J y_j, \quad y_1, \cdots, y_J \in \mathbb{R}, \quad y_{J+1} \leq 0 \\ &\text{such that } \sum_{j=1}^J y_j \geq -y_{J+1}(\Phi(t) - \Phi(t - \theta_j^i)) - (1 - \Phi(t - \theta_j^i))\theta_j^i \quad \forall (j,i). \end{aligned} \quad (53)$$

By weak duality, we have that

$$(52) \leq (53).$$

Therefore, to prove that Equation (51) is bounded from above by zero, it is sufficient to prove that its dual is bounded from above by zero.

Upper bound on the dual's objective function. To compute the upper bound on the dual's objective, it sufficient to observe that the dual's objective depends on values $(y_1, \cdots, y_J)$, which can be arbitrary in $\mathbb{R}$, but whose sum is constrained from above by the constraint in Equation (53). As a result, we have

$$\begin{aligned} (53) &\leq \min_{y_{J+1} \leq 0} \max_{\theta_j^i} -y_{J+1}(\Phi(t) - \Phi(t - \theta_j^i)) - (1 - \Phi(t - \theta_j^i))\theta_j^i \\ &= \min_{y \geq 0} \max_{\theta_j^i} y(\Phi(t) - \Phi(t - \theta_j^i)) - (1 - \Phi(t - \theta_j^i))\theta_j^i \\ &\leq \min_{y \geq 0} \max_{\check{\theta} \in [-1/\sqrt{\min_j \Sigma_{j,j}}, 1/\sqrt{\min_j \Sigma_{j,j}}]} y(\Phi(t) - \Phi(t - \check{\theta})) - (1 - \Phi(t - \check{\theta}))\check{\theta}. \end{aligned}$$

[36] The conditions in the above reference are satisfied since $\Phi(t - \theta_j)$ is a continuous function in θ_j, Θ is a compact space, and we can find a distribution so that the constraint holds with strict inequality.

The first equality follows directly by construction of the optimization program. The second equality is a change of variable where we wrote $-y_{J+1} = y$, and the third inequality substitutes the maximum over the set of (unknown) $n \times J$ parameters θ_j^i over a minimization over a parameter $\check{\theta}$ taking arbitrary values in the parameter space.

Claim of maximin optimality. To show maximin optimality it sufficies to show that the function

$$\inf_{y\geq 0} \sup_{\check{\theta}\in[-1/\sqrt{\min_j \Sigma_{j,j}},1/\sqrt{\min_j \Sigma_{j,j}}]} f(t,y,\check{\theta}), \quad f(t,y,\check{\theta}) := y(\Phi(t)-\Phi(t-\check{\theta})) - (1-\Phi(t-\check{\theta}))\check{\theta}$$

is bounded from above from zero for all t. Positivity can be shown numerically. We provide an analytical argument below.

Check the function value. Define

$$\check{\theta}(y) \in \arg \max_{\theta\in[-1/\sqrt{\min_j \Sigma_{j,j}},1/\sqrt{\min_j \Sigma_{j,j}}]} y(\Phi(t)-\Phi(t-\check{\theta})) - (1-\Phi(t-\check{\theta}))\check{\theta},$$
$$y^* \in \arg\min_{y\geq 0} y(\Phi(t)-\Phi(t-\check{\theta}(y))) - (1-\Phi(t-\check{\theta}(y)))\check{\theta}(y).$$

Suppose first that $\check{\theta}(\tilde{y}) = 0$ for some $\tilde{y} \geq 0$. Then it follows that

$$\min_{y\geq 0} f(t,y,\check{\theta}(y)) \leq f(t,\tilde{y},\check{\theta}(\tilde{y})) = 0.$$

We are left to discuss the case where $\check{\theta}(y) \neq 0$. Notice that $\frac{(1-\Phi(t-\check{\theta}(y))\check{\theta}(y)}{\Phi(t)-\Phi(t-\check{\theta}(y))} \geq 0$, for all $\check{\theta}(y) \neq 0$ since if $\check{\theta}(y) < 0$, the denominator and numerator are both negative and viceversa are both positive if $\check{\theta}(y) > 0$. Therefore, (assuming $\check{\theta}(y) \neq 0$, for all $y \geq 0$) we can always find a value $\tilde{y} \geq 0$ (since we minimize over $\tilde{y} \geq 0$), such that $\tilde{y} \leq \frac{(1-\Phi(t-\check{\theta}(\tilde{y}))\check{\theta}(\tilde{y})}{\Phi(t)-\Phi(t-\check{\theta}(\tilde{y}))}$. In such a case, we obtain a (weakly) negative valued objective function. As a result,

$$\min_{y\geq 0} y(\Phi(t)-\Phi(t-\check{\theta}(y))) - (1-\Phi(t-\check{\theta}(y)))\check{\theta}(y) \leq 0.$$

Admissibility. Admissibility follows directly from most-powerful claimed in Proposition 4, and Proposition 3, where, in this case, admissibility is with respect to a point-mass distribution over some $\theta \in \Theta_1(\epsilon)$, for some small enough $\epsilon > 0$.

B.2.9 Proof of Proposition 10

The proof mimics the proof of Proposition 1, and uses the assumption of unconstrained $\mathcal{R}$. Observe that since $\Theta_0^y \neq \emptyset$, there exist a $\theta \in \Theta_0^y$ such that $v_r(\theta) \leq 0$ for any $r \in \mathcal{R}$. Then the

maximin utility equals zero, since otherwise the editor can set $r(X) = 0$ and achieve zero utility. Observe that the editor payoffs is negative for all $\theta \in \Theta_0^y$. Under Assumption 3, the editor's payoffs equal exactly zero for all $\theta \in \Theta_0^y$, as long as Equation (35) is satisfied. In addition, the regulator payoff's is always non-negative for $\theta \in \Theta \backslash \Theta_0^y$ and any r, for any $\theta \backslash \Theta_0^y$ each $u^g(\theta)$ is weakly positive. As a result if Equation (9) holds, r^* is maximin. Consider now the case where $\beta_r(\theta) > 0$ for some $\theta \in \Theta_0^y$. Then, we can find a $\theta \in \Theta_0^y$ which leads to negative utility, for any decision $r(\cdot)$. Assumption 3 is invoked for the equality in Equation (9) to be a weak inequality.

B.3 Corollaries

Here we collect the proofs of the corollaries that are not immediate.

B.3.1 Proof of Corollary 1

Observe that since

$$\sum_k \widetilde{\delta}_k(r(X)) \leq 1, \quad \widetilde{\delta}_k : \{0,1\}^J \mapsto \{0,1\}^{2^J-1},$$

we have that

$$P\Big(\widetilde{\delta}_k(r^*(X)) = 1 \text{ for some } k | \theta = 0\Big) = \sum_k P\Big(\widetilde{\delta}_k(r^*(X)) = 1 | \theta = 0\Big),$$

since the events are disjoint. The result directly follows from Equation (27).

B.3.2 Proof of Corollary 2

First, observe that we can write:

$$P\Big(\min\{X_1, \cdots, X_G\} > t | \theta\Big) = \prod_{g=1}^{G} P(X_g > t | \theta). \tag{54}$$

Observe now that maximin optimality of r^* follows from the fact that the recommendation function is decreasing in each θ_j and that for any $\theta \in \Theta_0^y$ the probability of discovery cannot exceed the cost. To observe consider the extreme case where $\theta_g \to \infty$ for each $g \in \{1, \cdots, G-1\}$ and $\Theta^G = 0$. Then $P(X_g > t | \theta = \infty) = 1$ while $P(X_G > t | \theta = 0) = C$. Hence $\prod_{g=1}^{G} P(X^g > t | \theta) = C$ satisfying maximin optimality. The same argument applies if we shuffle the indexes. Finally, we discuss the second part of the claim. To show lack of maximin optimality we choose $\theta^{1:(G-1)}$ large enough so that $P(X_g > t | \theta) = 1$ for all $g < G$. As a result, any allocation with $P(X_G > t | \theta) > C$ is not maximin optimal. The same result applies to any other coordinate $g \neq G$.

B.3.3 Proof of Corollary 3

We first prove the first statement (notice: we let $\gamma = 1$ for notational convenience).

Case with $C(J)/\gamma = \alpha$. We write

$$\sum_{k\in\{\kappa,\cdots,J\}} \binom{J}{k}(p^*)^k \leq \sum_{k\in\{0,\cdots,J\}} \binom{J}{k}(p^*)^k$$

By the Binomial theorem, we have

$$(1+p^*)^J = \sum_{k\in\{0,\cdots,J\}} \binom{J}{k}(p^*)^k.$$

We then observe that

$$(1+p^*)^J \leq \exp(p^* J).$$

Observe that for $p^* \leq 1/J$ the expression is $O(1)$. Therefore, any order $1/J$ or slower, guarantees that the publication probability is bounded from below by a finite constant. Faster order are instead not possible, since if p^* was of order faster than $1/J$ this would imply that the publication probability converges to zero. However, this would lead to a contradiction since the publication probability must equal $\alpha > 0$. To rule out orders of convergence slower than $1/J$, we use a contradiction argument. The argument works as follows: any order of convergence slower than $1/J$ for p^* implies that the publication probability converges to infinity. This would lead to a contradiction since the publication probability must equal $\alpha < 1$. First, take

$$\sum_{k\in\{\kappa,\cdots,J\}} \binom{J}{k}(p^*)^k \geq \binom{J}{\kappa}(p^*)^\kappa. \tag{55}$$

Suppose now that p^* is of order slower than $1/J$, e.g. $p^* \asymp h_J/(J)$, for some arbitrary $h_J \to \infty$ as $J \to \infty$. Then we have

$$\binom{J}{\kappa}(p^*)^\kappa \asymp (J-\kappa)^\kappa \frac{h_J^\kappa}{J^\kappa} \to \infty,$$

since $\kappa < \infty$, leading to a contradiction.

Case with $C(J)/\gamma = \alpha/J$. We first start from the lower bound. We observe that we can write

$$J\times \sum_{k\in\{\kappa,\cdots,J\}} \binom{J}{k}(p^*)^k \geq J\times \binom{J}{\kappa}(p^*)^\kappa \geq J\times \frac{(J-\kappa)^\kappa}{\kappa^\kappa}\frac{1}{J^{\kappa+1}} = \frac{(J-\kappa)^\kappa}{\kappa^\kappa}\frac{1}{J^\kappa} = \frac{1}{\kappa^\kappa}(1-\kappa/J)^\kappa > 0$$

hence bounded away from zero. We now move to the upper bound. Consider first the case where $\kappa = 1$. Using the binomial theorem, we can write

$$\sum_{k\in\{\kappa,\cdots,J\}} \binom{J}{k}(p^*)^k = (1+p^*)^J - \sum_{k\in\{0,\cdots,\kappa-1\}} \binom{J}{k}(p^*)^k = (1+p^*)^J - 1 \leq \exp(1/J) - 1.$$

where the second equality follows from the fact that $\kappa = 1$ and the last equality by the fact that for $\kappa = 1, p^* = 1/J^2$. Using the mean value theorem, we have

$$\exp(1/J) - 1 = \frac{1}{J} + O(\frac{1}{J^2}) = O(1/J)$$

completing the claim for $\kappa = 1$. Let $\kappa > 1$. We have

$$J \times \sum_{k\in\{\kappa,\cdots,J\}} \binom{J}{k}(p^*)^k \leq J \times \sum_{k\in\{\kappa,\cdots,J\}} \frac{J^k}{k(k-1)} \frac{1}{J^{(1+1/\kappa)k}} = \sum_{k\in\{\kappa,\cdots,J\}} \frac{1}{k(k-1)} \frac{J}{J^{k/\kappa}}.$$

The inequality follows from the fact that $\binom{J}{k} \leq \frac{J^k}{k(k-1)}$ for $k > 1$ by definition of the binomial coefficient. Observe that since the sum starts from $\kappa > 1$ we have that for each summand $\frac{J}{J^{k/\kappa}} \leq 1$. Hence, we can write

$$\sum_{k\in\{\kappa,\cdots,J\}} \frac{1}{k(k-1)} \frac{J}{J^{k/\kappa}} \leq \sum_{k\in\{\kappa,\cdots,J\}} \frac{1}{k(k-1)} = O(1)$$

completing the proof since $\kappa > 1$.