Pre-registration and private experimentation

Zion Little* Jacob Montgomery† Keith Schnakenberg‡

February 18, 2025

Abstract

This paper develops a game-theoretic model of scientific experimentation to examine how pre-registration – the practice of having researchers announce what analyses they plan to do before collecting data and reporting results – affects researcher behavior and publication outcomes in the absence of external enforcement mechanisms. We address two key questions about pre-registration: whether researchers will adhere to pre-registered protocols without commitment or formal enforcement and whether it improves the truth value of published claims. We show that researchers have an incentive to follow pre-registration protocols without external enforcement because rational belief updating by the audience leads to informal punishment of unexpected experiments. However, the practice of pre-registration does not necessarily improve the epistemic value of science: since many experimentation plan are credible, scientists can use pre-registration to announce a plan that maximizes their chances of publication, potentially increasing false positive rates.

PRELIMINARY AND INCOMPLETE. UP-TO-DATE VERSION HERE.

*Graduate Student in Political Science, Washington University in St. Louis
†Professor of Political Science, Washington University in St. Louis
‡Professor of Political Science, Washington University in St. Louis. keschnak@wustl.edu

Many social scientists are rightly interested in reforming process of producing and publishing empirical findings. The credibility of social scientific research is threatened by 'p-hacking', which occurs when researchers try out multiple specifications or empirical analyses and then selectively report those that produce significant results (Head et al., 2015). The practice of selective disclosure, along with reviewers' and editors' biases in favor of papers that reject null hypotheses, leads to the concern that many published results are false positives (Brodeur et al., 2023; Gerber and Malhotra, 2008) or not reproducible (Open Science Collaboration, 2015). One proposed remedy for the problem of p-hacking is pre-registration, a practice in which researchers declare what analyses they plan to do before seeing any experimental results (Gonzales and Cunningham, 2015; Nosek et al., 2018; Kupferschmidt, 2018). There is, however, considerable disagreement on how preregistration should be used and what the best practices should be (Bakker et al., 2020; Ikeda et al., 2019), reflecting a lack of a solid understanding of how the institution is supposed to work.

In this paper, we provide a model of scientific experimentation and communication of results in which a scientist has the opportunity to selectively disclose results to a reviewer but may announce an experimentation plan (pre-registration) ahead of time. We use the model to answer two questions about pre-registration. First, we ask: *Can pre-registration change behavior in the absense of enforcement or commitment power?* This question is important because of debates about the extent to which authors ought to be bound to pre-registration plans, either by journal policies or some other enforcement mechanism. To address this question we model pre-registration as pre-play cheap talk which does not tie the scientist's hands in any way during experimentation or reporting but which must occur prior to the scientist observing any experimental results. Our answer to the first question is yes, the scientist's pre-registration informs the audience about the plan she intends to follow and she acts as if bound by the plan. The reason pre-registration changes behavior without external enforcement is that informal enforcement occurs naturally through rational belief formation by the scientist's audience. Pre-registrations are credible in our model because, if the scientist expects her pre-registration plan to be believed, deviating from the plan and reporting an unexpected test leads the audience to believe that the expected but unreported tests probably produced results that were unfavorable to the scientist's goals.

The second question we ask is: *Does pre-registration reduce the publication of false positives cause by selective disclosure?* The answer coming from our model is no. In fact, since many pre-registered plans are credible and since the scientist has the opportunity to set the agenda by announcing a plan, pre-registration will tend to benefit the scientist. Since the scientist is motivated by publication, this may increase the

Tests/Successes	0 Successes	1 Success	2 Successes	3 Successes
0 Tests	$1/3$			
1 Test	$1/7$	$3/5$		
2 Tests	$1/19$	$1/3$	$9/11$	
3 Tests	$1/55$	$1/7$	$3/5$	$18/19$

Table 1: Posterior probabilities for sequences of tests from the motivating example under full transparency. Posterior beliefs at which Receiver is persuaded are colored green and posterior beliefs at which the reviewer is not persuaded are colored red.

publication of false positives. A key part of the intuition from this result is that *opportunities* for p-hacking can destroy the scientist's credibility since any positive result may hide a large number of negative ones. This means that *actual* p-hacking is ineffective and publications are rare. Pre-registration gives the scientist more credibility by leading to a better common understanding of her experimentation plan, which makes it easier for the scientist to convince her audience to publish her results, ironically making p-hacking more effective.

Both of the points above are illustrated by the following example.

Example 1. Consider a situation in which a researcher (Sender) must persuade a reviewer (Receiver) to accept a paper supporting a hypothesis, initially believed by both players to be true with probability $1/3$. The Receiver prefers to accept only if posterior probability of the hypothesis being true is at least $1/2$. The Sender chooses from multiple tests, each successful with probability $3/4$ if the hypothesis is true and $1/4$ if false, to validate the hypothesis. There are three tests, labeled 1, 2, and 3, which can be observed sequentially in any order. Each test incurs a cost $c > 0$, signifying time and effort. The Sender cannot falsify test results but may selectively report results. To emphasize the issue of p-hacking we assume c is close to zero for this example. The Sender aims to convince the Receiver, with a payoff of 1 for success and 0 otherwise. Table 1 shows the posterior belief after any known sequences of tests (i.e. what the posterior belief would be under full transparency rather than private experimentation). Notice that the Sender can always persuade the Receiver with two successful tests, but can persuade her with just one successful test if is sufficiently likely that only one test was performed.

In this game there are many equilibria which vary by the Receiver's knowledge of the Sender's test priorities, which impacts the feasibility of selective disclosure. In one equilibrium, tests occur in a fixed order – we call this *fully prioritized*. For example, if the Receiver expects the Sender to always start with test 1, then possibly do test 2, and finally test 3, the Sender will not deviate. If test 1 succeeds, the Sender

stops and reports it, persuading the Receiver. If test 1 fails, given a low cost c, the Sender proceeds to test 2, and stops if it fails. If test 2 succeeds, reporting it alone will not persuade the Receiver since it implies test 1 failed. But a successful test 3 after this can persuade the Receiver. Hence, the Sender persuades the Receiver with probabilities $57/64 \approx .89$ if the hypothesis is true and $19/64 \approx .29$ if false. The overall probability of persuasion is $\frac{1}{3}\frac{57}{64} + \frac{2}{3}\frac{19}{64} = \frac{95}{192} \approx .49$.

At the other extreme, there is an equilibrum in which tests might be conducted in any order (we call this *unprioritized*). In this equilibrium, the Sender at every stage either stops experimenting or chooses randomly from the set of tests not already conducted. However, this creates a credibility problem severe enough that the Sender must always produce two successful tests in order to persuade the Receiver. If the first test is successful, Sender will keep experimenting by selecting another test at random from the remaining two. If the next test is successful, she can stop and reveal two successful tests which will persuade the Receiver. If not, given c close to zero, she will perform one last test the outcome of which will determine whether or not the Receiver is persuaded. If the first test is unsucessful, Sender will keep experimenting until the next failure or until she obtains two successes. In this equilibrium, Sender persuades Receiver with probability $54/64 \approx .84$ when the hypothesis is true and with probability $\frac{10}{64} \approx .15$ when the hypothesis is false. Thus, the total probability of persuading the Receiver is $\frac{1}{3}\frac{54}{64} + \frac{2}{3}\frac{10}{64} = \frac{37}{96} \approx .38$.

The Sender can do better than both equilibria described above by playing an equilibrium in which Receiver has partial information about Sender's priorities (we call this *partially prioritized*). For instance, suppose Receiver thinks that tests 1 and 2 can occur in any order, but test 3 will only be conducted after tests 1 and 2. Sender will also not deviate from this equilibrium, and this partial prioritization is enough to persuade the Receiver following one positive test from either 1 or 2. Under this strategy, the probability of that Sender reveals test 1 was successful is $\frac{1}{2}\frac{3}{4} + \frac{1}{2}\frac{1}{4}\frac{3}{4} = \frac{15}{32}$ when the hypothesis is true and $\frac{1}{2}\frac{1}{4} + \frac{1}{2}\frac{3}{4}\frac{1}{4} = \frac{7}{32}$ when the hypothesis is false. Therefore the Receiver's posterior belief is

$$\frac{\frac{1}{3}\frac{15}{32}}{\frac{1}{3}\frac{15}{32} + \frac{2}{3}\frac{7}{32}} = \frac{15}{29} > \frac{1}{2}$$

and likewise if test 2 is revealed. It follows that Sender always stops before conducting Test 3: if either Test 1 or Test 2 are successful then no further tests are needed, and if neither Test 1 nor Test 2 are successful then it is no longer possible to persuade the Receiver. Thus, in the equilibrium with partially informative priorities, the Sender persuades the Receiver when the hypothesis is true with probability $\frac{15}{16} \approx .93$ and when

the hypothesis is false the Sender persuades the Receiver with probability $\frac{7}{16} \approx .43$. Thus the total probability of persuading the Receiver is $\frac{1}{3}\frac{15}{16} + \frac{2}{3}\frac{7}{16} \approx .6$.

Equilibrium	Sender payoff	Receiver payoff
Fully prioritized	.49	.77
Unprioritized	.38	.85
Partially prioritized	.6	.69

Table 2: Approximate Sender and Receiver payoffs for equilibria in Example 1.

Table 2 summarizes the payoffs for both players in each of the equilibria described above. Notice that the Sender and Receiver have opposite rankings of equilibria. As we will argue, adding pre-play communication in the form of unrestricted pre-registration tends to benefit the Sender by allowing Sender to coordinate players on the equilibrium that maximizes her payoff. A few other observations are worth noting. First, the Sender is not necessarily better off maximizing opportunities for p-hacking since doing so might create severe enough credibility problems that Receiver is too difficult to persuade. Instead, Sender is better off in an equilibrium that limits p-hacking to some extent but makes it far more effective. Relatedly, the relationship between *opportunities* for p-hacking and actual false positives due to p-hacking is complicated in light of a rational Receiver.

Example 1 illustrates the point that the experimentation game possesses many equilibria. The reason is that beliefs about the tests that the Sender will conduct are to some degree self-enforcing – the Sender does not want to reveal unexpected tests because this causes the Receiver to believe that the expected tests must have failed. These equilibria also differ with respect to the opportunitites for p-hacking created by selective disclosure. This example also gives a preview of the role pre-registration plays in our analysis – pre-registration is pre-play communication that helps the Sender affect equilibrium selection. In fact, as we show, pre-registration will tend to select equilibria that are better for the Sender which are not the equilibria most consistent with the scientific objective of publishing true results. Our complete analysis of the selection of equilibria in this example is deferred to Example 2.

This paper contributes to a small theoretical literature on pre-registration in empirical science. More broadly, our work contributes to the literature on private experimentation. We build directly on Felgenhauer and Schulte (2014) who model strategic private experimentation.[1] Our main point of departure from this existing work is that in our model the tests that the sender can perform carry labels that distinguish them from

[1]Other useful papers with hidden testing, though with different applications in mind, include Herresthal (2022) and Shishkin (2021).

each other. Substantively, this applies to situations often encountered in empirical research. For instance, a researcher might have access to several similar dependent variables, and, though reviewers cannot be certain which analyses were run and not reported, the dependent variables are all distinguishable from one another. The seemingly subtle difference of labeling tests or experiments can be consequential for behavior in the game since the labels of the tests performed might be informative in some equilibria. Furthermore, this opens the door for credible pre-analysis plans since the sender can use them to create common expectations for which tests would be performed first.

Other related work on private experimentation is inspired by the Bayesian Persuasion framework of Kamenica and Gentzkow (2011), in which the sender controls the false-positive and false-negative rate of every test. In Felgenhauer and Loerke (2017), the Sender can choose a dynamic experimentation strategy – that is, they can design a whole dynamic path of experimentation contingent on each signal realization. They then reveal whatever subset of the experimental results they would like to reveal and conceal whatever they would like to conceal (though they cannot lie). Receivers would prefer to allow private experimentation and senders would not. This is because sender overcomes the lack of credibility associated with private experimentation by making the first signal to be so informative that it commits them to stopping after the first experiment. This is great for the receiver, who wants to have better information, but not so great for the sender, who must conduct a more precise experiment than she would like. Libgober (2022) applies a similar logic to an environment specifically tailored to the application of peer review. Relative to prior work, this adds a multidimensional signal structure that makes more explicit how the sender might substitute observed for unobserved actions to be more persuasive.[2] Our paper reaches similar conclusions about how limiting private experimentation benefits the Sender, though we prefer the model in which experiments have exogenous precision over the Bayesian persuasion approach. Our reasoning is that, though empirical researchers choose between various tests, they do not typically have precise control over the false-positive and false-negative rates of individual tests.[3]

Williams (2023) also writes on pre-registration. In that paper, pre-registration is useful because it signals private information acquired before any experimentation. In our model, the sender has no private information at the beginning of the game the pre-registration only functions as preplay communication to

[2]Felgenhauer and Xu (2021) uses the same basic structure as Felgenhauer and Loerke (2017) but the focus is on how the standard of evidence changes with or without transparency.

[3]Even when researchers are free to choose their sample size, doing so does not manipulate the different error probabilities independently. Furthermore, statistical power concerns are often related to resource constraints more than strategic concerns about persuasion.

communicate the sender's intentions.

1 A model of private experimentation with pre-registration

1.1 Game description

The players are a Sender and a Receiver. Both players are initially uninformed about the state of the world $\omega \in \{0,1\}$ with $\Pr[\omega = 1] = p \in (0,1)$ The Sender has access to an experimentation technology that generates signals about the state of the world. Specifically, Sender can perform any number of tests in the finite set $\mathscr{T} \coloneqq \{1,2,\ldots,T\}$. Each test t has a result $y_t \in \{0,1\}$, interpreted as failure (0) or success (1). The distribution of test results is $\Pr[y_t = 1|\omega = 1] = \lambda_1$ and $\Pr[y_t = 1|\omega = 0] = \lambda_0$. The tests are informative in the sense that $\lambda_1 > \lambda_0$. The Sender can perform the tests sequentially and at any time can choose to stop experimenting and reveal any subset of the tests and results. The Sender cannot falsify results. Thus, a report reveals (t, y_t) for some subset of the tests that the Sender conducted, where t tells the Receiver which test was performed and y_t gives the result of test t. The Receiver does not observe that a test was performed or the outcome of the test unless the Sender chooses to disclose it.

The order of play is as follows. First, the Sender chooses a cheap talk message m from some set $\mathscr{M}$. For now we will be purposely vague about the nature of $\mathscr{M}$ except that we will assume it is sufficiently rich to capture any information the Sender may want to convey about her strategy. Second, the Sender chooses whether to begin experimentation, which proceeds as we described above. Finally, the Receiver observes a report and chooses a.

The Receiver's utility is

$u_R(a,\omega)$	$\omega = 0$	$\omega = 1$
$a = 0$	1	0
$a = 1$	0	1,

The Sender's utility is $u_S(a) = a - cN$ where $c \geq 0$ is the cost of a test and N is the total number of tests performed. Notice that the Receiver should choose $a = 1$ only if the probability of $\omega = 1$ is greater than one half. We assume that $p < \frac{1}{2}$ which implies that the Receiver is initially inclined to choose $a = 0$.[4]

[4]It is straightforward to generalize the Receiver's preferences so that she prefers to choose $a = 1$ when the probability of $\omega = 1$ is greater than some threshold other than $\frac{1}{2}$. A threshold of $\frac{1}{2}$ is convenient and this added generality did not seem to produce significantly more insight.

1.2 Strategies and equilibrium concept

The Sender's strategy consists of a messaging strategy and an experimentation strategy σ_S. The Sender's messaging strategy is a pure strategy specifying a message in $\mathscr{M}$ to announce at the beginning of the game. To define the Sender's experimentation strategy, let h denote an experimentation history of the game and $\mathscr{H}$ the set of all possible experimentation histories. We can summarize the relevant aspects of an experimentation history as follows: Let $\mathscr{T}(h) \subseteq \mathscr{T}$ denote the tests that Sender has conducted at history h, and let $\eta(h) = \bigcup_{t \in \mathscr{T}(h)} \{(t, y_t)\}$ denote the set of tests and results from experimentation at history h. At each history h, the Sender's strategy σ_S specifies: (1) whether or not to continue experimenting, (2) a probability distribution over what test to perform next if the Sender continues experimenting, (3) a report $r \subseteq \eta(h)$ to send to the Receiver if the Sender does not continue experimenting. The Sender's beliefs at each experimentation history are denoted by $\mu_S(h) \coloneqq \Pr[\omega = 1|h]$. The Receiver observes a report R and her strategy, denoted $a_R(r)$, assigns a decision of either $a = 0$ or $a = 1$ to every possible report r from the Sender. The Receiver's beliefs following a report R are $\mu_R(r) \coloneqq \Pr[\omega = 1|r]$.

We will characterize perfect Bayesian equilibria (PBE) of the experimentation game and, to analyze the effect of pre-registration, apply an additional refinement motivated by preplay communication. A PBE to the experimentation game is an assessment such that each player's strategy is sequentially rational at every information set, μ_S is consistent with Bayesian updating given the observed experiments, and μ_R is consistent with Bayesian updating given the Sender's strategy and the observed experiments where possible. A PBE is selected by preplay communication if there is no credible message that the Sender would prefer to send: we define the relevant concepts in detail in Section 4.

2 Receiver's decision

The Receiver's optimal decision rule is to choose $a = 1$ if $\mu_R(r) > 1/2$ and $a = 0$ if $\mu_R(r) < 1/2$. For the sake of completeness, we assume that the Receiver also chooses $a = 1$ when she is indifferent, though this does not affect the analysis.

If the Receiver perfectly observed the experimentation history, then at a history h with exactly g successes and b failures then her beliefs would be

$$\Pr[\omega = 1] \coloneqq \pi(h) = \frac{p\lambda_1^g(1-\lambda_1)^b}{p\lambda_1^g(1-\lambda_1)^b + (1-p)\lambda_0^g(1-\lambda_0)^b}. \tag{1}$$

To compute Receiver's beliefs we must marginalize over possible histories consistent with a report r. Given a report r, the probability that the true experimentation history was h is $\Pr[h|r,\sigma_S] = \frac{\Pr[r\wedge h|n|\sigma_S]}{\Pr[r|\sigma_S]}$, so Receiver's beliefs following a report r are

$$\mu_R(r) = \sum_{h\in\mathscr{H}} \frac{\Pr[r\wedge h|n|\sigma_S]}{\Pr[r|\sigma_S]}\pi(h). \tag{2}$$

Clearly $\Pr[h|r,\sigma_S] = 0$ if $r \not\subset h$, so the intuition behind (2) is that the Receiver considers all of the unreported tests that Sender may have conducted.

Priorities The Sender's experimentation strategy contains two types of information. First, the Sender's strategy contains information about the order in which tests should be conducted when the Sender keeps experimenting. We call this information about *priorities*, which we will define more precisely below. Second, given fixed priorities, the Sender's strategy tells us at which histories she would stop experimenting. To analyze these two pieces of information separately, we next introduce a notion of *priorities*, capturing the order in which the Sender chooses experiments.

Definition 1. Let $W_{\mathscr{T}}$ denote the set of total preorders on $\mathscr{T}$ and let $P_{\mathscr{T}} \subset W_{\mathscr{T}}$ denote the set of strict orders on $\mathscr{T}$.[5] The Sender's *priorities* are defined as follows:

(a) The Sender's strategy is consistent with a relation $\succeq \in W_{\mathscr{T}}$ if, for any tests t and $t' \in \mathscr{T}$, $t \succeq t'$ implies that the Sender never conducts test t' at any history in which she has not conducted test t.

(b) *(Behavioral definition)* A relation $\succeq \in W_{\mathscr{T}}$ is a priority for the Sender if the Sender's strategy is consistent with $\succeq$ with probability one.

(c) *(Support-based definition)* Equivalently, a set $P' \subset P_{\mathscr{T}}$ of strict orders on $\mathscr{T}$ is a priority for the Sender if the Sender's strategy is consistent with some element of $P_{\mathscr{T}}$ with probability one.

Though the definitions above are equivalent, we retain both and use them interchangeably in the remainder of the paper. Our behavioral definition most naturally describes the Sender's behavioral strategy: at any given history, if the Sender continues experimenting, the selection of the next test is a probability distribution over the tests that have not yet been conducted and that do not have a strictly lower priority than any remaining tests. The support-based definition instead describes the support of the probability distribution over strict orders that is induced by the Sender's mixed strategy. This is perhaps less intuitive but useful for

[5] That is, $W_{\mathscr{T}}$ is the set of binary relations on $\mathscr{T}$ satisfying reflexivity, transitivity, and completeness. $P_{\mathscr{T}}$ contains all elements of $W_{\mathscr{T}}$ that also satisfy antisymmetry.

our analytical results. We use standard notation for the strict and weak parts of a binary relation, which we define below for the sake of completeness.

Definition 2. For any $\succeq \in W_{\mathscr{T}}$, we have:

1. $t \sim t'$ if and only if $t \succeq t'$ and $t' \succeq t$. If $t \sim t'$ we say these tests have the same priority under $\succeq$.

2. $t \succ t'$ if and only if $t \succeq t'$ and we do not have not $t' \succeq t$. If $t \succ t'$ we say t' has a strictly lower priority than t under $\succeq$.

We offer several remarks about the interpretation of the Sender's priorities before moving on to our main results.

Remark 1. Definition 1 gives two equivalent definitions of priorities which we use interchangeably in the remainder of the paper. The equivalence between the two definitions is established as follows: Let σ_S be consistent with the relation $\succeq \in W_{\mathscr{T}}$. Define $P(\succeq) = \{\succ \in P_{\mathscr{T}} : t \succeq t' \Rightarrow t \succ t' \forall t, t' \in \mathscr{T}\}$. Then σ_S is consistent with some element of P with probability one. In the other direction, if σ_S is consistent with some element of P with probability one then we never have t' conducted before t when $t \succeq t'$, so it must also be the case that σ_S is consistent with $\succeq$.

Remark 2. Priorities only constrain the Sender's behavior when she continues experimenting but provide no information about when the Sender should stop experimenting. For instance, a strategy in which the Sender stops experimenting immediately is consistent with any priorities. In general, any strategy in which some tests are conducted with probability zero is consistent with a class of priorities that differ with respect to the rankings of tests that are never conducted.

Remark 3. The equilibria described in Example 1 are easily described in terms of the Sender's priorities. The *unprioritized* equilibrium from Example 1 is consistent with Sender having priorities $\succeq_U$ defined by $1 \sim_U 2 \sim_U 3$ (where $t \sim t'$ indicates that t and t' are in an equivalence class, i.e. $t \succeq t'$ and $t' \succeq t$) using the behavioral definition or by $P = P_{\mathscr{T}}$ using the support-based definition. That is, all tests are in the same equivalence class and all strict orders are allowed under the unprioritized strategy. The fully prioritized equilibrium is consistent with Sender having priorities $\succeq_F$ defined by $1 \succ_F 2 \succ_F 3$ according to the behavioral definition and $P = \succeq_U$ using the support-based definition. That is, σ_S must be consistent with a particular strict order in which test 3 is only conducted after tests 1 and 2 and test 2 is only conducted after

test 1. Finally, the partiall prioritized equilibrium in Example 1 is consistent with priorities $\succeq_{Pa}$ such that $1 \sim_{Pa} 2 \succ_{Pa} 3$ using the behavioral definition or $P_{Pa} = \{\succ \in P_{\mathscr{T}} : 1 \succ 3 \ \& \ 2 \succ 3\}$ using the support-based definition.

3 Equilibrium experimentation

Having introduced the idea of the Sender's priorities, we are now ready to characterize equilibria to the experimentation game. As we state in Proposition 1 below, the relationship between priorities and equilibria is straightforward: there exist equilibria consistent with all possible priorities. As in Example 1, the reason for the existence of equilibria associated with any priorities is that "skipping ahead" in the expected order causes the Receiver to believe that any test that was expected to be conducted but not disclosed must have failed. Therefore, informal enforcement of priorities arises naturally from Bayeisan updating by the Receiver.

Proposition 1. *Let $\succeq \in W_{\mathscr{T}}$ be a particular but arbitrarily selected total preorder on $\mathscr{T}$. There exists an equilibrium in which:*

a.) Sender's experimentation behavior is consistent with the priorities $\succeq$.

b.) At any history, the Sender either stops experimenting or randomizes uniformly among the remaining tests that have the highest priority.

The proof of Proposition 1 makes use of the equivalence of Definitions 1(b) and 1(c) of the Sender's priorities, which mirrors the relationship between behavioral strategies and mixed strategies in sequential games. We prove the result in two steps. First, we consider particular but arbitrarily selected priorities $\succeq$ and consider an alternative game in which an order of tests $\succ$ is chosen by Nature uniformly from $P(\succeq)$, the set of strict orders that do not violate $\succeq$ as defined in Remark 1. This is a finite game and possesses an equilibrium, which is defined only by the Sender's stopping rule and the Receiver's decision rule. Let $\hat{v}(\succ, P(\succeq))$ denote the Sender's equilibrium payoff when the order of tests is $\succ$ and both players believe that the order of tests are drawn uniformly from $P(\succeq)$. The second step is to show that $\hat{v}(\succ, P(\succeq)) = \hat{v}(\succ', P(\succeq))$ for all $\succ, \succ' \in P(\succeq)$ and $\hat{v}(\succ, P(\succeq)) \geq \hat{v}(\succ'', P(\succeq))$ for all $\succ \in P(\succeq)$ and $\succ'' \neq \in P(\succeq)$. Thus, the Sender is indifferent over all orders consistent with the priorities and weakly prefers them to all orders inconsistent with priorities. This implies that, when we return to the original game and replace the move by Nature with randomization by the Sender, it is still incentive compatible for the Sender to randomize uniformly over elements of $P(\succeq)$.

Proposition 1 shows that Sender and Receiver could play many substantively different equilibria to the experimentation stage of our game. These equilibria differ with respect to how much information the Receiver has about the order in which Sender would perform the tests. As a result, the same experimentation game may possess equilibria in which selective disclosure is not important to the outcomes of the game and others in which selective disclosure produces substantial p-hacking or destroys the credibility of the Sender. For instance, there is always an equilibrium consistent with a plan that strictly orders all tests as in the "fully prioritized" equilibrium in Example 1. In this equilibrium, there is no uncertainty created by selective disclosure since the Receiver always knows what tests were conducted after any report. Thus, the experimentation game possesses equilibria with no p-hacking. At the other extreme, there is always an equilibrium consistent with all tests having the same priority as in the "unprioritized" equilibrium in Example 1, in which Receiver is uncertain which tests were conducted.

As we have already seen in Example 1, the Sender may benefit from playing an equilibrium with some strict priorities in order to enhance her credibility. Proposition 2 shows that, when the temptation to p-hack is the strongest (i.e. when c is small) and when the opportunity to p-hack is the greatest (i.e. when T is large), the Sender always benefits from playing an equilibrium consistent with some strict priorities.

Proposition 2. *There exist $c^* > 0$ and $T^* \in \mathscr{N}$ such that $c \leq c^*$ and $T \geq T^*$ implies that a Sender-preferred equilibrium strictly prioritizes some tests over others.*

The Sender sometimes prefers to coordinate on equilibria involving more informative plans in order to enhance her credibility. With a native or non-strategic Receiver, the Sender may prefer to maximize her opportunities for p-hacking and selective disclosure. With a sophisticated Receiver, however, opportunities for p-hacking reduce the value of any particular set of disclosed tests since the Receiver knows that the Sender might conceal a number of failed tests. Thus, coordinating on equilibria with more informative plans might make the Sender better off by increasing her credibility. When the opportunities for selective disclosure are in principle the highest, i.e. when there are a large number of available tests and conducting more tests is not very costly for the Sender, the Sender will always prefer to commit to some informative priorities. In fact, persuading the Receiver may be impossible under an unprioritized equilibria but reasonably likely under the optimal plan.

4 Pre-registration

We have discussed the multiplicity of equilibria to the experimentation game and given some motivation for why the Sender may want to communicate informative plans to the Receiver. To complete our argument we need to show why some equilibria may be more likely than others to be selected when there is pre-registration.

We treat pre-registration as cheap talk that must occur prior to any experimentation. Since the Sender does not have any private information when she announces a plan, pre-registration in our model is a form of pre-play communication or, as Farrell and Rabin (1996) put it, communication about intentions. That is, pre-registration informs the Receiver not about experimental results or about the state of the world, but about the experimentation plan that the Sender intends to follow.

We argue that pre-registration in this game will tend to select equilibria that improve the Sender's ex ante expected payoff. However, with no further assumptions, the addition of pre-play communication does not reduce the set of equilibria at all. For instance, for any equilibrium to the game without communication, there is an equilibrium with pre-play cheap talk in which the Sender chooses an arbitrary message, the Receiver ignores the message, and the players choose strategies consistent with that equilibrium. However, prior work shows that this result stems from the often unrealistic assumption in cheap talk models that messages have no meaning other than how players endogenously interpret them in equilibrium. If the players share a common language and the game satisfies certain credibility properties, then pre-play communication may favor certain equilibria over others. In particular, when one player has a monopoly on pre-play communication, these conditions favor equilibria that maximize that player's payoff. Thus, we will show that our game satisfies the relevant credibility conditions with respect to the Sender's experimentation plans.

We first define the set of messages that the Sender may use. Proposition 1 suggests that priorities are useful messages since there is a straightforward mapping from priorities to PBE of the experimentation game. Thus, we assume that the Sender communicates about priorities in pre-registration:

Assumption 1. The set of messages available to the Sender is $\mathscr{M} = 2^{P_{\mathscr{T}}}$. That is, the Sender can announce any priorities as operationalized by Definition 1(c).

To explain how pre-registration helps us understand equilibrium selection in the experimentation game, we lean on prior work on preplay communication (e.g. Farrell, 1988; Aumann, 1990; Farrell and Rabin, 1996; Lo, 2021). The departure of muich of this work from standard cheap talk models is the assumption

the players share a common language. As Farrell (1993) points out, in contrast to standard cheap talk models in game theory in which messages have no intrinsic meaning and are given a substantive interpretation only in equilibrium, most communication occurs in settings in which the messages have a literal interpretation. For example, a pre-registration announcement $m \in \mathscr{M}$ can be literally intepreted as a statement that the Sender does not intend to perform the tests in any order that conflicts with these stated priorities. When this is true, the Receiver has more guidance on how to interpret an unexpected message: the Receiver can understand the literal meaning of the message and then simply ask herself whether that message is credible.

What makes an unexpected message credible? One criterion proposed by Farrell (1988) is that the message be *self-committing* in the sense that, if Sender thinks that her message will be believed, she is better off acting consistent with the message. Aumann (1990), however, argues that this criterion is insufficient because the Sender may be better off having Receiver believe some messages even if she does intend to follow them. Therefore, according to Aumann, a credible message must also be *self-signaling* in the sense that the Sender *only* wants a message to be believed if she intends to follow it. Our equilibrium selection criterion captures the idea that an equilibrium is implausible if there is a credible (i.e. self-signaling) message that the Sender would prefer to send in the pre-registration phase. The justification, as in prior work, is to imagine that the pre-registration announcements have a literal interpretation, so that the Receiver knows their intrinsic meaning and asks only whether the message is credible enough to be believed.

We now define the concept self-signaling as it relates to our game. Since our interest is in refining equilibria rather than justifying Nash equilibrium itself, we define these credibility conditions relative to a particular equilibrium to be tested, as in Farrell (1993) and Matthews, Okuno-Fujiwara and Postlewaite (1991). For any priorities $\succeq$ represented by a set of strict orders $P(\succeq)$ and for any particular strict order $\succ \in P_{\mathscr{T}}$, let $\hat{v}(\succ, P(\succeq))$ denote the Sender's payoff from the order of tests being $\succ$ when Receiver believes they are drawn uniformly from $P(\succeq)$. We define our notion of crediblity below.

Definition 3. A message $m \in \mathscr{M}$ is credible relative to an equilibrium with priorities $\succeq$ if and only if:

(a) For all $\succ'' \in P(\succeq)$ and $\succ' \in P(\succeq) \backslash m$, $\hat{v}(\succ', m) < \hat{v}(\succ'', P(\succeq))$; and

(b) For all $\succ'' \in P(\succeq)$ and $\succ' \in m \backslash P(\succeq)$, $\hat{v}(\succ'', b(m)) < \hat{v}(\succ'', b(P(\succeq)))$.

Definition 3 translates the idea of self-signaling (Baliga and Morris, 2002; Lo, 2021) into the language of our model. Both parts of Definition 3 represent the idea that Sender would only want to send the message

m if she intended to follow it, but consider two different relationships between the orders in the message m and the putative equilibrium to be tested. Definition b (a) applies when the message m *excludes* some order of tests $\succ'$ that was part of the original equilibrium. In this case, m is only credible if, when the Receiver believes the message m but the Sender uses the order $\succ'$ anyway, the Sender is worse off than in the original equilibrium. Definition b (b) applies when the message m includes some strict orders $\succ'$ that were not part of the original equilibrium. In this case, the message m is only credible if the Sender is made worse off when the Receiver believes the message m but the Sender follows the original equilibrium.

We have argued that an equilibrium is implausible under pre-registration if there is a credible message that the Sender would rather send. Assumption 2 establishes this as our equilibrium selection criterion.

Assumption 2. The players will not play an equilibrium under pre-registration if there is a message that is credible and would make the Sender better off relative to that equilibrium.

Before moving on to the main result of this section, it is useful to revisit Example 1 to understand how our equilibrium selection criterion would affect that simpler game. We do this below in Example 2.

Example 2. To illustrate our theory of how pre-registration affects equilibrium selection, we return to Example 1. Recall that we characterized three equilibria to this game, which we labeled unprioritized, partially prioritized, and fully prioritized. Each equilibrium is associated with a message, which we label $m_{\text{Un}}, m_{\text{Pa}}$, and m_F for the unprioritized, partially prioritized, and fully prioritized equilibria respectively. For any $i, j, k \in \{1,2,3\}$ let $\succ_{ijk}$ be the strict order for which $i \succ_{ijk} j \succ_{ijk} k$. Therefore we have:

$$m_{\text{Un}} = \{\succ_{123}, \succ_{213}, \succ_{132}, \succ_{231}, \succ_{312}, \succ_{321}\} \tag{3}$$

$$m_{\text{Pa}} = \{\succ_{123}, \succ_{213}\} \tag{4}$$

$$m_F = \{\succ_{123}\}. \tag{5}$$

Below, we test whether each of these equilibria should be eliminated under pre-registration according to our criteria.

- *Unprioritized equilibrium.* Recall that this equilibrium gives the Sender the lowest payoff, so the Sender prefers to announce m_{Pa} or m_F if such an announcement is credible. Consider an announcement of m_{Pa}. Since $m_{\text{Pa}} \subset m_{Un}$, only part (a) of Definition 3 applies. This requires that the Sender's payoff from convincing the Receiver that she intends to play the partially prioritized strategy when

actually playing the unprioritized strategy is worse than her payoff in the unprioritized equilibrium. This is true of the message m_{Pa}: each order of tests that is allowed under m_{Pa} but not under m_{Un} moves test 3 first or second. However, since the partially prioritized equilibrium has test 3 conducted last, a Receiver who observes test 3 will only ch oose $a = 1$ if *both* of tests 1 and 2 also succeed. Thus, the Sender would not want the Receiver to believe that she was playing the partially prioritized strategy if she was playing the unprioritized strategy. Thus, m_{Pa} is a credible message that should persaude the Receiver that the Sender actually intends to play the partially prioritized strategy. This implies that the unprioritized equilibrium should be eliminated under pre-registration.

- *Partially prioritized equilibrium.* Assumption 2 requires not only that a suggestion to play another equilibrium is credible but that the Sender actually prefers to play the alternative equilibrium. Since the partially prioritized equilibrium is sender-preferred, it is never eliminated under pre-registration.

- *Fully prioritized equilibrium.* The Sender prefers to play the partially prioritized equilibrium over the fully prioritized equilibrium. Hence, the relevant question is whether m_{Pa} is credible. Since $m_F \subset m_{\text{Pa}}$, the relevant condition is part (b) of Definition 3. This requires that the Sender's payoff from convincing the Receiver that she plans to play the partially prioritized strategy when she actually plans to play the fully prioritized strategy is strictly worse than her payoff in the fully prioritized equilibrium. This condition fails: in fact, since $\succ_{123}$ is permissible under both strategies, the payoff to announcing m_{Pa} but only playing the strategy consistent with $\succ_{123}$ is equal to the payoff from the partially prioritized equilibrium, which is the Sender-preferred equilibrium. Thus, the Sender would want the Receiver to believe that she is playing the partially prioritized equilibrium rather than the fully prioritized strategy even if that is not true. Thus, m_{Pa} is not credible and preregistration does not eliminate the fully prioritized equilibrium.

Hence, pre-registration eliminates the unprioritized equilibrium but does not eliminate the fully or partially prioritized equilibria.

In Example 2 we can rule out the most vague plan that is also bad for the Sender. As it turns out, this is illustrative of a broader result: a proposal to play a more restrictive experimentation plan is credible because the Receiver will informally punish the disclosure of unexpected tests. Therefore, pre-registration eliminates an equilibrium if there is a more informative plan that is better for the Sender.

Proposition 3. *Let Assumption 1 and Assumption 2 hold. Then a PBE in which Sender's priorities are $\succeq$ is eliminated by pre-registration if there is an announcement m such that (a) the Sender prefers the equilibrium in which her priorities are m to the one in which her priorities are $\succeq$, and (b) $m \subset P(\succeq)$.*

Proposition 3 paints the picture the pre-registration may tend to select plans that are more specific and better for the Sender relative to the equilibrium that would have been played absent pre-registration. However, since any statement about the effects of pre-registration depends on a conjecture about what equilibrium would have been played without pre-registration, we can only make weak statements about the effect of the institution.

One way to interpret the results is to think of common conjectures about experimentation plans as arising from social conventions within fields or subfields or smaller scientific communities. In some scientific communities there may be strong conventions about which tests are most appropriate and therefore the most natural equilibrium absent any pre-registration might be something more closely approximating the fully prioritized equilibrium of Examples 1 and 2. In these communities, we may expect pre-registration to have no effect – reviewers already have firm ideas about what tests will be conducted and more vague plans that would benefit the Sender are not credible. In other scientific communities, we may think that there are no strong conventions at all. In those situations we may think that the natural equilibrium is something more closely approximating the unprioritized equilibrium of Examples 1 and 2. In these cases, pre-registration will strengthen common expectations about which tests will be conducted. However, the pre-registration plans are chosen to benefit the Sender and therefore will tend to increase the rate at which false positives are published.

5 Discussion and conclusions

In our game of private experimentation, we have shown that pre-registration is credible without enforcement but that it benefits the Sender rather than decreasing the rate at which false positives are published. In fact, pre-registration in our model may increase the effectiveness of p-hacking by increasing the Sender's credibility.

One question raised by our analysis is whether adding institutional enforcement may improve outcomes. One way to improve outcomes using institutional enforcement would be to use an all-or-nothing rule for experimentation plans: for each priority level, the Sender is required either to report every test at that priority level or to not report any of them. This counteracts the Sender's strategy of calibrating selective disclosure

to balance credibility with opportunities for selective disclosure. That said, this essentially amounts to requiring full disclosure and if we could do that we would not have much use for pre-registration.

Our analysis also omits several issues with real world pre-registration that we think are amenable to formal modeling and are interesting directions for future work. For example, reviewers may not always read pre-analysis plans. This might undercut the self-committing nature of these plans, though a full equilibrium analysis is needed to understand these effects.

References

Aumann, Robert. 1990. "Nash equilibria are not self-enforcing." *Economic decision making: Games, econometrics and optimisation* pp. 201–206.

Bakker, Marjan, Coosje L. S. Veldkamp, Marcel A. L. M. van Assen, Elise A. V. Crompvoets, How Hwee Ong, Brian A. Nosek, Courtney K. Soderberg, David Mellor and Jelte M. Wicherts. 2020. "Ensuring the quality and specificity of preregistrations." *PLOS Biology* 18(12):1–18.
URL: *https://doi.org/10.1371/journal.pbio.3000937*

Baliga, Sandeep and Stephen Morris. 2002. "Co-ordination, spillovers, and cheap talk." *Journal of Economic Theory* 105(2):450–468.

Brodeur, Abel, Scott Carrell, David Figlio and Lester Lusher. 2023. "Unpacking P-hacking and Publication Bias." *American Economic Review* 113(11):2974–3002.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20210795*

Farrell, Joseph. 1988. "Communication, coordination and Nash equilibrium." *Economics Letters* 27(3):209–214.
URL: *https://www.sciencedirect.com/science/article/pii/0165176588901723*

Farrell, Joseph. 1993. "Meaning and Credibility in Cheap-Talk Games." *Games and Economic Behavior* 5(4):514–531.
URL: *https://www.sciencedirect.com/science/article/pii/S0899825683710298*

Farrell, Joseph and Matthew Rabin. 1996. "Cheap Talk." *Journal of Economic Perspectives* 10(3):103–118.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.10.3.103*

Felgenhauer, Mike and Elisabeth Schulte. 2014. "Strategic Private Experimentation." *American Economic Journal: Microeconomics* 6(4):74–105.
URL: *http://www.jstor.org/stable/43189689*

Felgenhauer, Mike and Fangya Xu. 2021. "THE FACE VALUE OF ARGUMENTS WITH AND WITHOUT MANIPULATION." *International Economic Review* 62(1):277–293.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12479*

Felgenhauer, Mike and Petra Loerke. 2017. "BAYESIAN PERSUASION WITH PRIVATE EXPERIMENTATION." *International Economic Review* 58(3):829–855.
URL: *http://www.jstor.org/stable/45018773*

Gerber, Alan and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3(3):313–326.
URL: *http://dx.doi.org/10.1561/100.00008024*

Gonzales, Joseph E and Corbin A Cunningham. 2015. "The promise of pre-registration in psychological research." *Psychological Science Agenda* 29(8):2014–2017.

Head, Megan L, Luke Holman, Rob Lanfear, Andrew T Kahn and Michael D Jennions. 2015. "The extent and consequences of p-hacking in science." *PLoS biology* 13(3):e1002106.

Herresthal, Claudia. 2022. "Hidden testing and selective disclosure of evidence." *Journal of Economic Theory* 200:105402.
URL: *https://www.sciencedirect.com/science/article/pii/S0022053121002192*

Ikeda, Ayumi, Haoqin Xu, Naoto Fuji, Siqi Zhu and Yuki Yamada. 2019. "Questionable research practices following pre-registration." *Japanese Psychological Review* 62(3):281–295.

Kamenica, Emir and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101(6):2590–2615.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590*

Kupferschmidt, Kai. 2018. "A recipe for rigor." *Science* 361(6408):1192–1193.
URL: *https://www.science.org/doi/abs/10.1126/science.361.6408.1192*

Libgober, Jonathan. 2022. "False Positives and Transparency." *American Economic Journal: Microeconomics* 14(2):478–505.

URL: *https://www.aeaweb.org/articles?id=10.1257/mic.20190218*

Lo, Melody. 2021. "Language and coordination games." *Economic Theory* 72(1):49–92.

URL: *https://doi.org/10.1007/s00199-020-01279-9*

Matthews, Steven A, Masahiro Okuno-Fujiwara and Andrew Postlewaite. 1991. "Refining cheap-talk equilibria." *Journal of Economic Theory* 55(2):247–273.

URL: *https://www.sciencedirect.com/science/article/pii/002205319190040B*

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven and David T. Mellor. 2018. "The preregistration revolution." *Proceedings of the National Academy of Sciences* 115(11):2600–2606.

URL: *https://www.pnas.org/doi/abs/10.1073/pnas.1708274114*

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251):aac4716.

URL: *https://www.science.org/doi/abs/10.1126/science.aac4716*

Shishkin, Denis. 2021. Evidence Acquisition and Voluntary Disclosure. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21 New York, NY, USA: Association for Computing Machinery p. 817–818.

URL: *https://doi.org/10.1145/3465456.3467586*

Williams, Cole Randall. 2023. "Preregistration and Incentives." *SSRN Electronic Journal* .

URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract*$_i d = 3796813$