Strategic Sample Selection[*]

Alfredo Di Tillio[†] Marco Ottaviani[‡] Peter Norman Sørensen[§]

July 15, 2018

Abstract

What is the impact of sample selection on the inference payoff of an evaluator facing a monotone decision problem? We show that anticipated selection increases or decreases the accuracy of a statistical experiment according to whether the reverse hazard rate of the data distribution is log-supermodular—as in location experiments with normal noise—or log-submodular. The results are applied to the analysis of strategic sample selection by a biased researcher and extended to the case of uncertain and unanticipated selection. Our theoretical analysis offers applied research a new angle on the problem of selection in empirical and experimental studies, by characterizing when sample selectivity, selective assignment to treatment, and strategic omission of variables benefit or hurt the evaluator.

Keywords: Strategic selection; Comparison of experiments; Accuracy; Persuasion; Welfare

JEL codes: D82, D83, C72, C90

[*] Research funded by the European Research Council through Advanced Grant 29583 (EVALIDEA). We thank Matteo Camboni for outstanding research assistance. We also thank, without implication, Pierpaolo Battigalli, Francesco Corielli, Nicola Limodio, and seminar participants at Bocconi, Bruxelles, Caltech, Carlos III, Chicago, Como, Duke, East Anglia, EUI, Genova, Gerzensee, Graz, Helsinki, Innsbruck, Leicester, Lisbon, London Business School, Milan, Northwestern, Nottingham, Paris, Rome, Southampton, Surrey, Tel Aviv, Toulouse, and Vienna for helpful comments.

[†] Department of Economics and IGIER, Bocconi University, Via Roberto Sarfatti 25, 20136 Milan, Italy. Phone: +39–02–5836–5422. E-mail: alfredo.ditillio@unibocconi.it.

[‡] Department of Economics and IGIER, Bocconi University, Via Roberto Sarfatti 25, 20136 Milan, Italy. Phone: +39–02–5836–3385. E-mail: marco.ottaviani@unibocconi.it.

[§] Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK–1353 Copenhagen K, Denmark. Phone: +45–3532–3056. E-mail: peter.sorensen@econ.ku.dk.

1 Introduction

Observational data are often nonrandomly selected, due to choices made by the subjects under investigation or sample inclusion decisions by data analysts.[1] Suppose a new treatment is taken by the healthiest patients rather than by random patients in a group. Because of selection, positive outcomes become more likely, but they provide weaker evidence that the treatment is effective. Balancing these two effects, how does sample selection affect the quality of inference? Is the evaluator's assessment of the treatment effect more accurate with selective or with random assignment? When estimating a regression coefficient of interest, is a selected sample more or less informative than a random sample with the same sample size? In a different, yet similar vein, is a regression with a random missing covariate more or less informative than a regression with a strategically omitted variable?

Experimental data can also suffer from selection problems challenging internal validity, when researchers subvert the random allocation of subjects to treatment rather than control.[2] Similar questions arise in a number of other contexts. For instance, the right of peremptory challenge gives a defendant the option to strike down a number of jurors. Given that the defendant selects the most favorable jurors, how is the quality of final judgement affected? When feeding consumer reviews to potential buyers with limited attention, should an e-commerce platform post random reviews or allow the merchant to cherry-pick them? When testing a student in an exam, should the teacher ask questions at random or allow the student to select the most preferred questions out of a batch?

These comparisons are all instances of one and the same issue: assessing the impact of selection in a monotone decision problem. There is an unknown real-valued state θ, e.g. the effect of a new treatment, or the true value of a regression coefficient. An evaluator must choose an action, e.g. approve the new treatment, or estimate the coefficient, knowing that marginally increasing the action decreases the payoff in low states and increases it in high states. The evaluator decides after observing the realization of a statistical experiment consisting of a random vector $X = (X_1, \dots, X_n)$. For instance, in a location problem, observations have the form $x_i = \theta + \varepsilon_i$, where ε_i represents the baseline health of an individual, or the noise term in a regression.[3] Our main question is, in which of the following scenarios does the evaluator make better decisions:

[1] For instance, from the outset Heckman (1979) refers to these two sources of selection.

[2] See, for example, Schulz (1995) and Berger (2005) for extensive accounts of subversion of randomization in clinical trials.

[3] In other applications, θ may represent a defendant's level of guilt, the quality of a merchant's good, or a student's ability, and ε_i a juror's bias, a reviewer's leniency, or a student's specific knowledge of a certain topic.

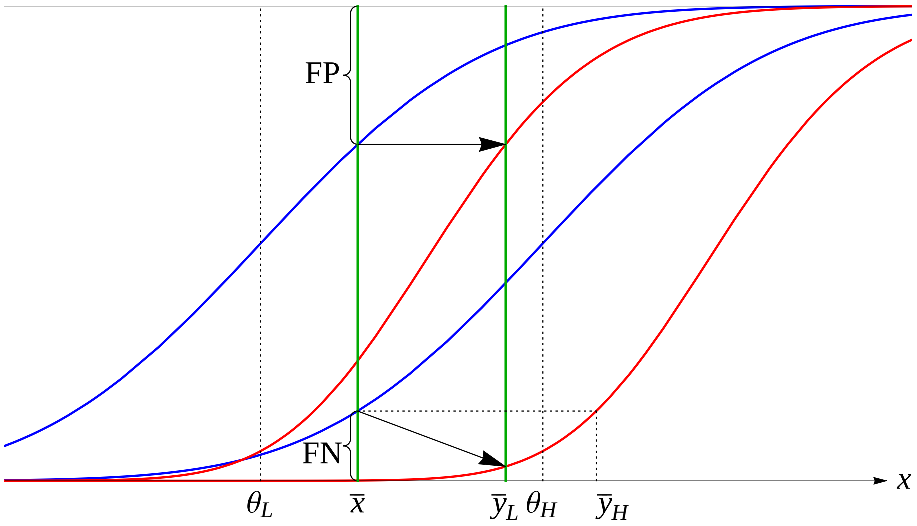

Figure 1: Simple hypothesis testing with a normal experiment: selection provides more accurate information, decreasing false negatives while keeping false positives the same.

- **Random Experiment**. The n observations are i.i.d. draws from a state-dependent cumulative distribution $F(\cdot|\theta)$. In a location problem, this means that the noise terms $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. draws from a known cumulative distribution $F(\cdot)$, and $F(x|\theta)=F(x-\theta)$.

- **Selected Experiment**. The n observations are selected—possibly strategically, by another party—as the n highest out of $k>n$ presampled i.i.d. draws from $F(\cdot|\theta)$. In a location problem, equivalently, $\varepsilon_1,\ldots,\varepsilon_n$ are the n highest of $k\geqslant n$ i.i.d. draws from F.

Following a standard approach pioneered by Blackwell (1951, 1953), we address this question by comparing the probability distributions over actions that the evaluator can induce, in each state, in the two experiments. The approach is easy to illustrate in the special case of a simple hypothesis testing problem: two states, a low state θ_L and a high state θ_H, and two actions, rejection—the correct choice in the low state—and acceptance—the correct choice in the high state. In this case the evaluator's decision is the familiar trade-off between the probability of a false positive—accepting in the low state—and the probability of a false negative—rejecting in the high state.

Consider a one-dimensional (sample size $n=1$) location problem with a normal noise distribution F. In the random experiment, the evaluator observes the realization of a random variable that is normally distributed with mean θ_L in the low state, and mean θ_H in the high state, as illustrated by the blue curves in Figure 1. The evaluator resolves the trade-off between false positives and false negatives by accepting if and only if the observed realization x exceeds some cutoff point $\bar{x}$.[4]

[4] A location problem with normal noise, like every other experiment considered in this paper, satisfies the monotone

The optimally chosen probability of a false positive is then $1-F(\bar{x}-\theta_L)$, denoted by FP in the figure, while the probability of a false negative is $F(\bar{x}-\theta_H)$, denoted by FN.

How does a selected experiment with presample size $k>1$ compare? In the selected experiment, the noise distribution becomes F^k, that is, the evaluator observes a random variable distributed according to $F^k(x-\theta_L)$ in the low state, and $F^k(x-\theta_H)$ in the high state—these distributions are represented by the red curves in Figure 1. In this example with normal noise, the selected experiment turns out to be better: by adopting the (possibly suboptimal) cutoff point $\bar{y}_L$ in this experiment, the evaluator induces as many false positives, because $1-F^k(\bar{y}_L-\theta_L)=1-F(\bar{x}-\theta_L)$, but also induces fewer false negatives, because $F^k(\bar{y}_L-\theta_H)<F(\bar{x}-\theta_H)$.

What explains the beneficial impact of selection just described? To answer this question we start from an observation made by Lehmann (1988), who pointed out that an equivalent way to formulate the property that $\bar{y}_L$ induces as many false positives and fewer false negatives is to say that the selected experiment is *more accurate*.[5] This means that the cutoff point $\bar{y}_H$ that induces as many false *negatives*, defined by the equation $F^k(\bar{y}_H-\theta_H)=F(\bar{x}-\theta_H)$, is larger than $\bar{y}_L$.[6] Indeed, by adopting the smaller cutoff $\bar{y}_L$ the evaluator necessarily induces more acceptance—and in particular more acceptance in the high state—than by adopting the larger cutoff $\bar{y}_H$. Note that this is exactly what happens in the normal case depicted in Figure 1.

Our first main result identifies a necessary and sufficient condition for a larger presample size k to increase or decrease accuracy in one-dimensional location experiments. Theorem 1 shows that accuracy monotonically increases in presample size k if and only if the reverse hazard function of the basic noise distribution, $-\log F$, is logconcave, as with normal or logistic noise. Likewise, accuracy monotonically decreases in k if and only if $-\log F$ is logconvex, as with exponential noise. Intuitively, our logconcavity criterion requires that neither the top tail of the distribution should be thicker than in the Gumbel distribution, nor the bottom tail should be thinner—for otherwise the thickening of the upper tail or the thinning of the bottom tail created by selection would decrease accuracy. For example, the condition implies that the evaluator always gains from sample selection when F is normal or logistic, but always loses when F is exponential.

To assess the impact of selection in the general case, with possibly non-additive noise and any

likelihood ratio property: given any two states, the higher the realization x, the higher the relative odds of the higher state. This property implies that the evaluator's optimal decision is increasing in x. With two actions and sample size $n=1$, this simply means choosing the higher action (acceptance) if and only if x is at least as large as some cutoff $\bar{x}$.

[5] The latter nomenclature is due to Persico (2000).

[6] Going beyond binary state, and comparing arbitrary experiments X and Y with distributions $F(\cdot|\theta)$ and $G(\cdot|\theta)$, the property that Y is more accurate than X means that the function $(G(\cdot|\theta))^{-1}(F(x|\theta))$ is increasing in θ for every x.

sample size $n \geqslant 1$, we employ a natural generalization of Lehmann's (1988) concept of accuracy, which we call *conditional accuracy*. This notion is new to our paper, and allows comparisons between any pair of experiments, with arbitrary patterns of correlation among observations. Conditional accuracy shares the basic intuition with (and, for $n = 1$, it reduces to) Lehmann's (1988) original notion. To get this intuition in the most transparent way, consider the simple hypothesis testing setup. In an n-dimensional experiment X the evaluator again adopts a cutoff strategy, but now the cutoff is a more complicated object, an $(n-1)$-dimensional curve in $\mathbb{R}^n$. For example, with a vector of i.i.d. observations $x = (x_1, \ldots, x_n)$ from a location experiment with normal noise, the evaluator accepts if and only if the average observation exceeds a certain threshold.[7] We say that another experiment Y is conditionally more accurate than X if the (suitably defined) cutoff curve that induces as many false positives in Y as in X, lies entirely below the (analogously defined) curve that induces as many false negatives. By Lehmann's (1988) argument we can then show that in Y the evaluator can achieve as many false positives as in X, but fewer false negatives.

The notion of conditional accuracy is the key technical tool needed to tackle the new issues arising in the multidimensional case. Indeed, the main difficulty in comparing multidimensional selected experiments is precisely due to the selected observations $x_1, \ldots, x_n$ being correlated with each other, even conditionally on the state θ. Our tool allows us to disentangle the net value of information added by each observation, and hence understand when selection adds or subtracts value to the evaluator's problem as the presample size k increases. Our second and most important result, Theorem 2, shows that conditional accuracy monotonically increases or decreases in presample size k, according to whether the reverse hazard rate $f(x|\theta)/F(x|\theta)$ is log-supermodular or log-submodular. In a location experiment, log-supermodularity reduces to logconcavity of the noise distribution's reverse hazard rate f/F. This condition strengthens the logconcavity criterion in Theorem 1 by adding a regularity condition. Intuitively, the noise distribution must be *increasingly* thinner at the top and thicker at the bottom, compared to the Gumbel distribution.

Our results have important implications for applied research. While typically thought of exclusively as a threat to internal validity, selective assignment based on untreated outcomes—or, more generally, some unobservable characteristics correlated with untreated outcomes—can actually benefit an evaluator who properly anticipates selection. Actually, as we discuss at the end of the paper, selection may benefit even an unwary evaluator who does not anticipate it. Similar considerations apply in a regression context. Whereas other forms of selection—as occurring, for instance, in truncated regression—decrease accuracy, selection based on order statistics is beneficial (for a fixed sample size) under the conditions identified in our theorems. By the same token, a

[7] As is well known, in the normal case the average observation is a sufficient statistic for the whole vector of observations. In this case, the cutoff curve has the form $\sum_i x_i/n = \tilde{x}$ for some $\tilde{x}$.

regression with a random missing covariate can be less informative than a regression with, say, a variable omitted in order to maximize selection bias.

Selection naturally arises when the evidence is provided by a strategic researcher who observes a presample $x_1,...,x_k$ of size k and then reports the n most favorable realizations to the evaluator. Given the presample size k and the fact that the evaluator uses an acceptance cutoff rule, it is indeed optimal for the researcher to report the highest realizations. We provide a strategic foundation for sample selection by introducing a researcher whose payoff increases in the evaluator's action. For example, the researcher aims at convincing the evaluator that the true state is high—e.g. that a new treatment is effective, or that a regression coefficient is large. The researcher's incentives to bias upward the evaluator's inference through sample selection are anticipated in equilibrium by the evaluator. Under the conditions in our theorems, equilibrium selective sampling benefits also the researcher in the empirically relevant case where the evaluator a priori favors rejection; when instead the evaluator a priori favors acceptance, equilibrium selection harms the researcher.

We also endogenize the amount of selection in terms of costly investment by the researcher in obtaining the k presample realizations. The evaluator's anticipation and resulting adjustment for selection partly frustrates the researcher's attempt to manipulate. If selection is fully anticipated in equilibrium—for example because the researcher's cost of presample collection is known—then sample selection is a pure rat race when the noise follows a Gumbel distribution. In that case, selection has no impact on the payoffs of the two parties. The result is a loss by the researcher exactly equal to the cost of presample collection. Thus, with Gumbel noise the researcher unambiguously benefits from commitment not to allow (or, equivalently, to disclose) presample collection.[8]

Finally, we illustrate how an evaluator who can ex ante forbid sample selection may prefer not to do so, in order to incentivize the researcher to provide more evidence. Intuitively, consider a situation in which the researcher benefits from selection because of the increased acceptance rate that results from selective reporting. Given that the researcher's individual rationality constraint is relaxed, the researcher's incentives to collect a presample larger than the required sample ends up benefitting the evaluator under our logconcavity condition on the reverse hazard rate. Through this mechanism, the evaluator can benefit from limiting the sample size, tolerating sample selection from a presample larger than the sample, and committing not to look at more data.

The final part of the paper discusses the welfare impact of unanticipated and uncertain selection. Surprisingly, there are natural situations—in particular, the realistic scenario in which the

[8] More generally, selection is not completely self defeating, even when the evaluator correctly anticipates the extent of selection k. Our results characterize the net impact of properly anticipated selection on acceptance probability (the researcher's decision payoff) and the evaluator's decision payoff.

evaluator a priori strongly favors rejection—in which selection benefits even when it is completely unanticipated. Intuitively, if the evaluator does not adjust for selection, acceptance occurs more often in every state, and the increase in false positives can be smaller than the reduction in false negatives. In addition we show that, while uncertainty about k tends to damage the evaluator,[9] a selected experiment with an uncertain k can still be better than a random experiment.

Related Literature. Concerns about data selection and manipulation have long been voiced by the science and medicine literature and have led to important policy responses.[10] However, there is a dearth of modeling in the area.[11] An early exception is Blackwell and Hodges (1957), who analyze how an evaluator should optimally design a sequential experiment to minimize *selection bias*, a term they coined to represent the fraction of times a strategic researcher is able to correctly forecast the treatment assignment.[12] However, they did not model the information available to the researcher at the assignment stage. Moreover, the ensuing literature focused on exogenous selection bias and on how to adjust for it, rather than on its strategic origin and its impact on the quality of inference, on which we focus. Once we explicitly model information, we characterize situations in which selection actually benefits the evaluator, contrary to what Blackwell and Hodges (1957) stipulate.

Relative to work on optimal persuasion following Rayo and Segal (2010) and Kamenica and Gentzkow (2011), in our setting information acquisition is costly and information manipulation is naturally constrained by the need of reporting a signal selected from the presample. With sample size $n = 1$, our researcher discloses a single observation, as in the limited-attention model first proposed by Fishman and Hagerty (1990).[13] Thus, we have a signal-jamming model of equilibrium

[9] This effect is most transparent in the Gumbel case, as indifference to selection hinges on the evaluator's exact knowledge of the extent of selection k.

[10] See, for example, the analysis of Schulz, Chalmers, Hayes, and Altman (1995) and the CONSORT statement, http://www.consort-statement.org.

[11] Glaeser (2008) discusses a number of issues in this regard. Di Tillio, Ottaviani, and Sørensen (2017) compare different types of selection in the context of an illustrative model with binary noise, which violates the logconcavity assumption maintained in this paper.

[12] Blackwell and Hodges (1957) argue that selection bias is minimized by a truncated binomial design, according to which the initial allocations to treatment and control are selected independently with a fair coin, until half of the subjects are allocated to either treatment or control; from that point on, allocation is deterministic. Efron (1971), instead, characterizes the selection bias resulting from a biased coin design, according to which the probability of current assignment to treatment is higher if previous randomizations resulted in excess balance of controls over treatments.

[13] See also Henry (2009), Dahm, Gonzàlez, and Porteiro (2009), Felgenhauer and Schulte (2014), Hoffmann, Inderst, and Ottaviani (2014), and Herresthal (2017) for persuasion models with endogenous information acquisition. Henry and Ottaviani (2015) analyze a dynamic model of persuasion with costly information acquisition à la Wald (1945), where information is truthfully reported at the time of application.

persuasion through presample collection and then sample selection. The researcher's choice of the size k of the presample is akin to the agent's effort choice in Holmström's (1999) classic career concern model. The wrinkle here is that this effort results in private information, which the researcher then uses to select the reported information.

In a complementary approach to modeling conflicts of interest in statistical testing, Banerjee, Chassang, Monteiro, and Snowberg (2017a) propose a theory of an ambiguity-averse researcher facing an adversarial evaluator.[14] In another complementary approach, Tetenov (2016) analyzes an evaluator's optimal commitment to a decision rule when privately informed researchers select into costly testing. Instead, we focus on the impact of a researcher's manipulation of data on the welfare of an uncommitted evaluator.

2 Statistical Setup

An evaluator chooses an action from a finite set $A = \{a_1, \dots, a_L\} \subseteq \mathbb{R}$ with $a_1 < \cdots < a_L$ under uncertainty about the true value of a state $\theta \in \Theta \subseteq \mathbb{R}$, where Θ is either a finite set or a (possibly unbounded) interval. The evaluator holds a prior belief on the state, represented by a density (or mass, if Θ is finite) function π, and a payoff function $u : \Theta \times A \to \mathbb{R}$ defining a *monotone decision problem*: there exist states $\theta_1 \leqslant \cdots \leqslant \theta_{L-1}$ such that, for every state θ and every $1 \leqslant \ell < L$, the difference $u(\theta, a_{\ell+1}) - u(\theta, a_\ell)$ is nonpositive if $\theta \leqslant \theta_\ell$ and nonnegative if $\theta > \theta_\ell$.

Information and Optimal Decision. Before deciding, the evaluator observes the realization of an *experiment*, an n-dimensional random vector $X = (X_1, \dots, X_n)$ taking values in some domain $D \subseteq \mathbb{R}^n$ and distributed according to a state-dependent density function $g(\cdot|\theta)$ satisfying the monotone likelihood ratio (MLR) property: for any two realizations $x, x' \in D$ with $x' \geqq x$, the ratio $g(x'|\theta)/g(x|\theta)$ is increasing in θ.[15] The MLR property has an important consequence in monotone problems: the evaluator can, without loss, limit attention to *monotone strategies*, where the chosen action increases with the observed realization of X.[16]

[14] See also Kasy (2016) and Banerjee, Chassang, and Snowberg (2017b).

[15] Here and in the remainder of the paper, given two vectors $x = (x_1, \dots, x_n)$ and $x' = (x'_1, \dots, x'_n)$, we say that x' is larger than x, and write $x' \geqq x$, to indicate that $x'_i \geqslant x_i$ for every $i = 1, \dots, n$. Similarly, we say x' is smaller, and write $x' \leqq x$, to indicate that $x'_i \leqslant x_i$ for every $i = 1, \dots, n$.

[16] By Bayes' rule, the MLR property implies that the evaluator's posterior belief on the state increases with the observed realization x in the *likelihood ratio order*, that is, for every x and $x' \geqq x$ the ratio $\pi(\theta|x')/\pi(\theta|x)$ is increasing in θ. Thus, by our assumption on payoffs, the evaluator cannot lose by (weakly) increasing the chosen action in response to a higher realization. For a proof of this claim, see Quah and Strulovici (2009, Theorem 2).

Simple Hypothesis Testing. The most elementary instance of a monotone decision problem—our leading example in much of our discussion below—is a *simple hypothesis testing* problem. There are two states, a low state θ_L and a high state $\theta_H > \theta_L$, and two actions, rejection (a_1) and acceptance (a_2). The evaluator would like to reject in the low state and accept in the high state, so that payoffs satisfy $u(\theta_L, a_2) \leqslant u(\theta_L, a_1)$ and $u(\theta_H, a_2) \geqslant u(\theta_H, a_1)$. Given a realization x, the evaluator optimally accepts if and only if

$$g(x|\theta_H) \,/\, g(x|\theta_L) \geqslant r, \tag{1}$$

where r depends on the problem parameters.[17] In a one-dimensional ($n = 1$) experiment, this monotone strategy takes a familiar form: accept if and only if $x \geqslant \bar{x}$, where $\bar{x}$ is the cutoff point satisfying (1) with equality. In general, with $n \geqslant 1$, the realizations x satisfying (1) with equality define a curve in $\mathbb{R}^n$, above which the evaluator accepts. More precisely, the acceptance region defined by (1) is an *upper set* in the experiment's domain, i.e. a set $U \subseteq D$ containing every point of D that is larger than some point of U. Given the decision to accept in U and reject in $D \setminus U$, it is easy to see that the evaluator's optimal expected payoff can be written (disregarding constants) as

$$-\, r \underbrace{\Pr_L(X \in U)}_{\text{prob. false positive}} - \underbrace{\Pr_H(X \notin U)}_{\text{prob. false negative}}\,, \tag{2}$$

a negatively weighted sum of the probability of a false positive (accepting in the low state) and the probability of a false negative (rejecting in the high state), with r serving as relative weight.

Random vs. Selected Experiments. Our main concern in this paper is the welfare comparison between experiments of a particular form: the evaluator observes the n highest of $k \geqslant n$ random variables that are conditionally i.i.d. given the state. Formally, given a family of distribution functions $(F(\cdot|\theta))_{\theta\in\Theta}$ with associated densities $(f(\cdot|\theta))_{\theta\in\Theta}$ satisfying the MLR property, a *selected experiment* is a random vector $X = (X_1, \dots, X_n)$ where, for each state θ, the random variable X_1 is the highest of $k \geqslant n$ random draws from $F(\cdot|\theta)$, the random variable X_2 the second highest, and so on. Thus, the random vector X takes values in the domain

$$D = \mathbb{R}^n_{\geqslant} := \{x \in \mathbb{R}^n : x_1 \geqslant \cdots \geqslant x_n\},$$

and the density function of X in state θ is given by the following formula:

$$g(x|\theta) = \frac{k!}{(k-n)!} F^{k-n}(x_n|\theta) f(x_1|\theta) \cdots f(x_n|\theta).$$[18]

[17] The conditional probability of θ_H given that $X = x$ equals $\pi(\theta_H)g(x|\theta_H)/[\pi(\theta_L)g(x|\theta_L) + \pi(\theta_H)g(x|\theta_H)]$, that is, $1/[(\pi(\theta_L)/\pi(\theta_H))(g(x|\theta_L)/g(x|\theta_H)) + 1]$. Thus, the expected payoff difference between acceptance and rejection is nonnegative if and only if $g(x|\theta_H)/g(x|\theta_L) \geqslant r := [\pi(\theta_L)/\pi(\theta_H)][u(\theta_L, a_1) - u(\theta_L, a_2)]/[u(\theta_H, a_2) - u(\theta_H, a_1)$.

[18] The MLR property holds for this density because log-supermodularity is preserved by integration (Athey, 2002).

Note that the cumulative distribution function of X_1 is $F^k(\cdot|\theta)$ and, for every $i = 2,\ldots,n$, the conditional cumulative distribution function of X_i given that $X_1 = x_1,\ldots,X_{i-1} = x_{i-1}$ depends only on the last of these conditions, and is given by

$$\frac{F^{k-i+1}(x_i|\theta)}{F^{k-i+1}(x_{i-1}|\theta)} \qquad (x_i \leqslant x_{i-1}). \tag{3}$$

We call n the *sample size* and k the *presample size* of the selected experiment. When $k = n$, we call the experiment *random*, because it is informationally equivalent to n i.i.d. draws from $F(\cdot|\theta)$.[19]

A particular case that often arises in applications is the case of a *location experiment*, where the evaluator knows the shape of the data distribution, but does not know where the distribution is located. Formally, the distributions in the family $(F(\cdot|\theta))_{\theta\in\Theta}$ are all shifted versions of one and the same cumulative distribution function F admitting a logconcave density function f, that is,

$$F(x|\theta) = F(x-\theta) \qquad \forall\theta\in\Theta,\ \forall x\in\mathbb{R}.$$

In this case we call F the *basic noise distribution*, because the random variables $X_1,\ldots,X_n$ in a selected location experiment X with sample size n and presample size k have the following form: $X_1 = \theta + \varepsilon_{k:k},\ldots,X_n = \theta + \varepsilon_{k-n+1:k}$, where the noise term $\varepsilon_{k:k}$ is the highest of $k \geqslant n$ noise terms $\varepsilon_1,\ldots,\varepsilon_k$ randomly drawn from F, the noise term $\varepsilon_{k-1:k}$ the second highest, and so on.

Comparing One-Dimensional Experiments by Accuracy. Our comparison between selected experiments with sample size $n = 1$ is based on the notion of accuracy, first investigated by Lehmann (1988).[20] Given two arbitrary one-dimensional experiments X and Y with respective distributions $F(\cdot|\theta)$ and $G(\cdot|\theta)$, we say that Y as *more accurate* than X if the function $\zeta_\theta(x) = (G(\cdot|\theta))^{-1}(F(x|\theta))$ is increasing in θ for every x. This monotonicity criterion is a necessary and sufficient condition for Y to be preferred to X in every monotone decision problem: if Y is more accurate, then the evaluator can induce a state-by-state more favorable distribution over actions than with X, as shown in Lehmann (1988, Theorem 5.1); see also Proposition 1 below for the special case with $n = 1$. For example, in a simple hypothesis testing problem, as discussed in the introduction and in our illustration of Theorem 1 in the next section, the cutoff $\bar{y}_L = \zeta_L(\bar{x})$ inducing in Y the same false positives as in X, is smaller than the cutoff $\bar{y}_H = \zeta_H(\bar{x})$ matching the false

[19] Clearly, knowing in advance that in every state the random variables $X_1,\ldots,X_n$ are sorted so that X_1 is the highest, X_2 the second highest, etc. is of no value for the evaluator.

[20] Accuracy can be defined as Blackwell's (1951,1953) sufficiency, restricted to monotone decision problems involving one-dimensional experiments satisfying the MLR property. This class of problems was first studied by Karlin and Rubin (1956). The term *accuracy* was introduced in Persico (2000) and later adopted by Quah and Strulovici (2009). For applications of the notion of accuracy to economic problems, see also Jewitt (2007).

negatives.[21] This implies that $\bar{y}_L$ induces fewer false negatives in Y than $\bar{x}$ does in X, and a fortiori, recalling (2), that the evaluator must prefer Y to X. Our analysis of multidimensional experiments in Section 3.2 builds on a natural generalization of accuracy, that we will call *conditional accuracy*.

3 Welfare Impact of Selection

In this section we characterize the families of distributions $(F(\cdot|\theta))_{\theta\in\Theta}$ such that selection has a monotone welfare impact: the larger the presample size, the better (or the worse) the selected experiment. We begin our analysis with the simple case of one-dimensional location experiments. Our characterization is tighter and simpler to illustrate in this case, and it will provide a useful starting point for introducing the characterization in the general case.

3.1 One-Dimensional Selected Location Experiments

A selected location experiment with sample size $n = 1$ and presample size $k \geqslant 1$ has the form $X = \theta + \varepsilon_{k:k}$ where $\varepsilon_{k:k} = \max\{\varepsilon_1, \dots, \varepsilon_k\}$ and $\varepsilon_1, \dots, \varepsilon_k$ are random draws from a basic noise distribution F with logconcave density f. Our first main result characterizes the basic noise distributions F for which the evaluator's optimal expected payoff is increasing, and those for which it is decreasing, in the presample size k.

Theorem 1. *Fixing the sample size to $n = 1$, an increase in the presample size makes a selected location experiment more (resp. less) accurate if and only if the reverse hazard function of the basic noise distribution, $-\log F(\varepsilon)$, is logconcave (resp. logconvex) in ε.*

To gain intuition for the role of logconcavity of the reverse hazard function in our characterization,[22] it is helpful to discuss the result in the context of simple hypothesis testing. Let $X = \theta + \varepsilon_{k:k}$ and $Y = \theta + \varepsilon_{m:m}$ be selected experiments with sample size $n = 1$ and different presample sizes, k and $m \neq k$, respectively. Let $\bar{x}$ denote any cutoff point that the evaluator may set in experiment X, accepting if and only if $X \geqslant \bar{x}$. Then Y guarantees a larger payoff than X if, for any such cutoff point, there is a corresponding cutoff point $\bar{y}_L$ that, in experiment Y, leads to as many false positives

[21] Here and in the sequel, to ease notation, we write ζ_L and ζ_H instead of the more cumbersome ζ_{θ_L} and ζ_{θ_H}. We adopt analogous notations for other objects as well.

[22] Marshall and Olkin (2007) define the hazard function as $\log F$. Since F ranges between zero and one, $\log F$ is necessarily negative. Our definition uses a minus sign, so that logconcavity of the function makes sense.

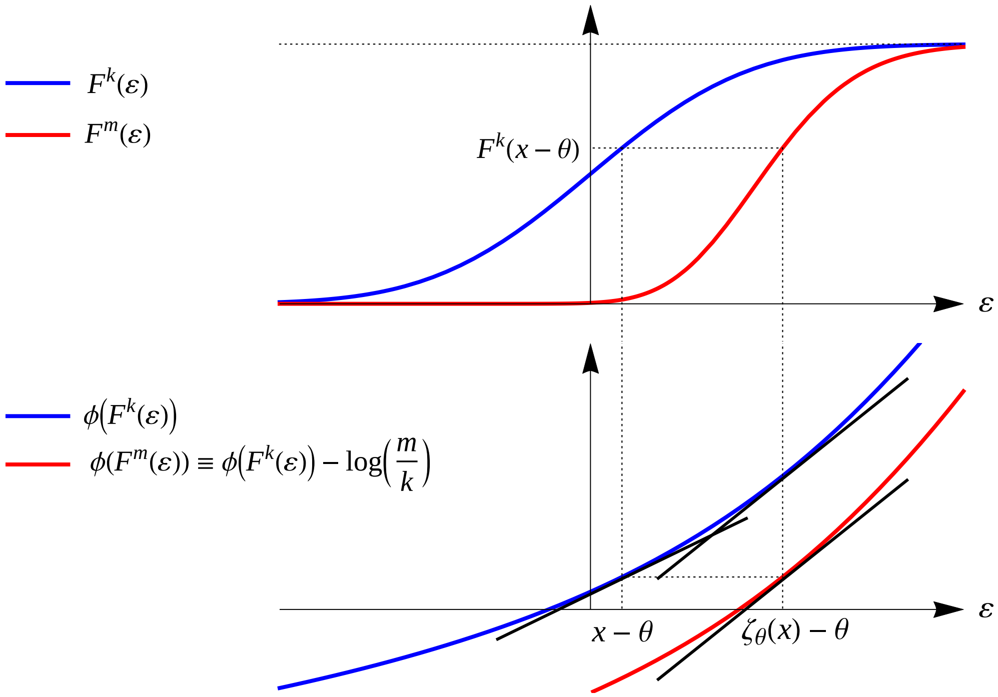

Figure 2: Normal one-dimensional location experiment: double-log transformation.

and fewer false negatives, so that

$$F^m(\bar{y}_L - \theta_L) = F^k(\bar{x} - \theta_L) \quad \text{and} \quad F^m(\bar{y}_L - \theta_H) \leqslant F^k(\bar{x} - \theta_H). \tag{4}$$

Solving the equation in (4) for $\bar{y}_L$ and plugging the result into the inequality in (4), we conclude that Y guarantees a higher payoff in every simple hypothesis testing problem (i.e. whatever values θ_L, θ_H and $\bar{x}$ may take) if and only if the function $\zeta_\theta(x) = (F^m)^{-1}\big(F^k(x-\theta)\big) + \theta$ is increasing in θ for every x, that is, if and only if Y is more accurate than X. Taking the derivative of $\zeta_\theta(x)$ with respect to θ, for every x we must then have

$$mF^{m-1}(\zeta_\theta(x) - \theta) f(\zeta_\theta(x) - \theta) \geqslant kF^{k-1}(x-\theta) f(x-\theta). \tag{5}$$

Equivalently, changing variable from x to $u = F^k(x-\theta)$, it must be the case that for all $u \in [0,1]$ the slope of F^m computed at the quantile $\zeta_\theta(x) - \theta = (F^m)^{-1}(u)$ is greater than the slope of F^k computed at the corresponding quantile $x - \theta = (F^k)^{-1}(u)$.

Under what conditions on F, m and k does property (5) hold for every x? To answer this question, it is convenient to first transform F^k and F^m in such a way that the transformed functions are parallel vertical shifts of each other. The suitable transformation is the strictly increasing function $u \mapsto \phi(u) = -\log(-\log u)$, because then $\phi(F^m(\cdot)) = \phi(F^k(\cdot)) - \log(m/k)$. The transformation is illustrated in Figure 2 for the case of a standard normal F and $m > k$.[23] Since our double-log transformation is strictly increasing, property (5) is equivalent to the slope of $\phi(F^k(\cdot))$ at $x-\theta$ being less than the slope of $\phi(F^m(\cdot))$ at $\zeta_\theta(x)-\theta$. Equivalently, the slope of $\phi(F^k(\cdot))$ itself is greater at $\zeta_\theta(x)-\theta$ than at $x-\theta$. Since $\phi(F^k(\cdot)) = \phi(F(\cdot)) - \log k$, this is satisfied when either $\phi(F(\cdot))$ is convex and $m > k$ (because $m > k$ implies $\zeta_\theta(x) \geqslant x$), or $\phi(F(\cdot))$ is concave and $m < k$ (because $m < k$ implies $\zeta_\theta(x) \leqslant x$). Thus, we conclude that a larger presample size benefits the evaluator when the reverse hazard function $-\log F$ is logconcave (as in the normal case illustrated in Figure 2) and hurts the evaluator when $-\log F$ is logconvex.

Dispersion and Accuracy. When X and Y are location experiments, so that $F(x|\theta) = F(x-\theta)$ and $G(y|\theta) = G(x-\theta)$ for some noise distributions F and G, we can also equivalently apply the notion of *dispersion*, instead of appealing to accuracy. Bickel and Lehmann (1979) define G as *less dispersed* than F if the quantile difference $G^{-1}(u) - F^{-1}(u)$ is decreasing in $u \in [0,1]$. This notion also appeared in our discussion (and is, in fact, used in the proof) of Theorem 1, when we asked whether the slope of F^m computed at $\zeta_\theta(x) - \theta = (F^m)^{-1}(u)$ is greater than the slope of F^k computed at $x-\theta = (F^k)^{-1}(u)$. Accuracy and dispersion are equivalent for location experiments. A location experiment is more accurate if and only if the noise distribution is less dispersed (Lehmann, 1988, Theorem 5.2).

Gumbel Distribution. There is only one distribution F such that $-\log F$ is both logconcave and logconvex, i.e. loglinear, namely the Gumbel (maximum) extreme value distribution, $F(\varepsilon) = \exp(-\exp(-\varepsilon))$. This distribution, which plays a special role in the ensuing analysis, is such that for every k the selected experiment with presample size k is neither less nor more accurate than a random experiment—the evaluator is indifferent to selection. The following intuitive argument also leads to the same conclusion. With a presample size equal to k, the noise distribution is $F^k(\varepsilon) = \exp(-k\exp(-\varepsilon)) = F(\varepsilon - \log k)$. Thus, compared to a random sample, selection inflates noise by a constant, $\log k$. The evaluator adjusts for this inflation, and is back to square one.

Logistic and (Generalized) Exponential Distributions. Besides the normal case discussed earlier, another instance where more selection is better for the evaluator is the logistic case, $F(\varepsilon) = 1/(1+e^{-\varepsilon})$. (We prove this and the following claim below.) Our main example of the opposite case, where more selection hurts the evaluator, is the exponential distribution $F(\varepsilon) = 1 - e^{-\varepsilon}$ (for $\varepsilon \geqslant 0$). More generally, given any $a < -1$, the distribution F such that

$$F(\varepsilon) = \exp\left(\frac{1}{1+a}\left[(1-\exp(-\varepsilon))^{1+a} - 1\right]\right) \qquad (\varepsilon > 0) \tag{6}$$

is such that $-\log F$ is logconvex. (The exponential distribution is the special case $a \to -1$.)

[23] The plots are drawn for $k = 1$ and $m = 8$, but any $k \geqslant 1$ and $m > k$ give the same qualitative result.

Contribution to Stochastic Ordering of Order Statistics. Previous results in the literature on stochastic ordering of order statistics only covered basic noise distributions with decreasing hazard rate. Notably, Khaledi and Kochar (2000, Theorem 2.1) showed that for any distribution with decreasing hazard rate higher order statistics are more dispersed.[24] Given that logconcavity implies increasing hazard rate by Prekopa's theorem, the only basic noise distribution with logconcave density for which Khaledi and Kochar's (2000) result applies is the exponential (loglinear) distribution, which has constant hazard rate.[25] The novel characterization in Theorem 1 applies more generally to the relevant case of basic noise distributions with logconcave densities.

Real Presample Size. According to our definition, the presample size is a natural number k, but the interpretation—as well as the statement in the theorem—for real numbers $k > 1$ is equally valid. Increasing selection from k to $m > k$ changes the noise distribution from F^k to $F^m = (F^k)^{m/k}$. This is akin to having basic noise distribution F^k and a fractional presample size $m/k > 1$. Our comparative statics result in Theorem 1 characterizes when this increases accuracy. Note the implication that the basic noise distribution F is such that selection monotonically benefits (or hurts) the evaluator if and only if the basic noise distribution F^k has the same property for every real number $k > 1$. Indeed, both properties are equivalent to logconcavity of $-\log F$.

3.2 General Multidimensional Selected Experiments

We now turn to the comparison between selected experiments from a (not necessarily location type) family of distributions $(F(\cdot|\theta))_{\theta\in\Theta}$ with any sample size $n \geqslant 1$. As we shall see, the results obtained in the context of one-dimensional location experiments for the common distributions discussed earlier (normal, Gumbel, exponential, etc.) carry over to arbitrary sample sizes. However, the extension of the characterization in Theorem 1 to the multidimensional case is far from immediate, and poses some important challenges.

In a selected experiment with sample size n and presample size k, the evaluator observes the highest, second highest, ..., nth highest of k random draws from $F(\cdot|\theta)$. Toward a generalization of our analysis beyond one-dimensional experiments, one may therefore try to grasp intuition from the individual comparisons of each intermediate order statistic with a random draw. However, following this approach would be misleading, for two reasons. First, even in those cases where the

[24] According to Khaledi and Kochar (2000, Theorem 2.1), if X_i's are i.i.d. with decreasing hazard rate, then $X_{i:n}$ is less dispersed than $X_{j:m}$ whenever $i \leqslant j$ and $n-i \geqslant m-j$. Setting $i = n = 1$ and $j = m = k$, we have that the maximum of k i.i.d. variables with decreasing hazard rate is more dispersed than the original variable.

[25] Theorem 1 also covers distributions with decreasing hazard rate, where $-\log F$ is necessarily logconvex.

n highest of $k > n$ realizations *are* better than, say, n random draws, an intermediate order statistic (say, the second highest, or third highest, etc.) is, in isolation, generally *not* more accurate than a random draw.[26] Second, as is evident from (3), the n highest realizations of $k \geqslant n$ random draws are correlated among themselves, even conditionally on the state. Intuitively, correlation tends to reduce information,[27] which creates further ambiguity about the sign of the net marginal value of information added by an intermediate order statistic.

To shed light on these issues, we now introduce a generalization of Lehmann's (1988) notion of accuracy, which allows the comparison between experiments (not necessarily selected experiments) featuring arbitrary correlation patterns.

Comparing Multidimensional Experiments by Conditional Accuracy. Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be experiments with respective domains D_X and D_Y. For every state θ, we define a function $\zeta_\theta : D_X \to D_Y$ such that in state θ the random vector $\zeta_\theta(X)$ has the same distribution as Y, as follows: $\zeta_\theta(x_1, \ldots, x_n) = (z_1, \ldots, z_n)$, where $z_1, \ldots, z_n$ are defined recursively by the equations below (for brevity, we write $< i$ for the indices $1, \ldots, i-1$):

$$\Pr_\theta(X_1 \leqslant x_1) = \Pr_\theta(Y_1 \leqslant z_1) \quad \text{and} \quad \Pr_\theta(X_i \leqslant x_i | X_{<i} = x_{<i}) = \Pr_\theta(Y_i \leqslant z_i | Y_{<i} = z_{<i}). \qquad (7)$$

Then we say that Y is *conditionally more accurate* than X if $\zeta_\theta(x)$ is an increasing function of θ for every $x \in D_X$. Note that if $n = 1$, the function ζ_θ is defined by the first equation in (7). In this case, our definition reduces to Lehmann's (1988).

In what sense is a conditionally more accurate experiment better for the evaluator? To gain intuition, consider simple hypothesis testing. In experiment X the evaluator optimally accepts if and only if $X \in U$, where U is some upper set. Now consider experiment Y and suppose that the evaluator, perhaps suboptimally, accepts if and only if $Y \in \zeta_L(U)$. How does this strategy

[26] Consider, for example, a location experiment where the basic noise distribution is the *positive* exponential distribution: $F(x|\theta) = F(x-\theta) = \exp(x-\theta)$ for $x \leqslant \theta$. Let $X_1, \ldots, X_k$ be i.i.d. draws from $F(x|\theta)$, and let Y_1 and Y_2 be the highest and the second highest of these draws. The cumulative distribution functions of Y_1 and Y_2 are given by $\Pr_\theta(Y_1 \leqslant y) = \exp(k(y-\theta))$ and $\Pr_\theta(Y_2 \leqslant y) = k\exp((k-1)(y-\theta)) - (k-1)\exp(k(y-\theta))$, respectively. Thus, the function ζ_θ such that $\zeta_\theta(X_1)$ has the same distribution as Y_1 is $\zeta_\theta(x) = (x-\theta)/k + \theta$, which increases with θ. Indeed, by Theorem 1, experiment Y_1 is more accurate than X_1, because $\log(-\log F(\varepsilon)) = \log(-\varepsilon + \theta)$ is (strictly) concave in ε. Now consider the function ζ'_θ such that $\zeta'_\theta(X_2)$ has the same distribution as Y_2. This function is U-shaped in θ, because its reciprocal, $\log(k\exp((k-1)(y-\theta)) - (k-1)\exp(k(y-\theta))) + \theta$, is a bell-shaped function of θ. Thus, Y_2 is neither less nor more accurate than X_2. Yet, the basic noise distribution $F(\cdot) = \exp(\cdot)$ satisfies the condition of Corollary 1 below, so the evaluator is better off with (Y_1, Y_2) than with (X_1, X_2) in every (monotone) problem.

[27] The statistical literature on the comparison of multidimensional experiments with correlated observations is rather small. Shaked and Tong (1990, 1993) identify conditions under which an experiment with correlated draws is less informative than an experiment with independent draws. Their results, however, do not apply to our context.

fare? In the low state it induces as many false positives, for $\Pr_L(Y \in \zeta_L(U)) = \Pr_L(X \in U)$ by the very definition of ζ_L. In the high state, it induces fewer false negatives: $\Pr_H(Y \in \zeta_L(U)) \geqslant \Pr_H(Y \in \zeta_H(U)) = \Pr_H(X \in U)$, where the inequality follows from U being an upper set and Y being conditionally more accurate than X, and the equality from the definition of ζ_H. Much like in Lehmann's (1988) argument, then, since Y can guarantee a state-by-state higher payoff than X, a fortiori it must give at least as much expected payoff.

The analogous argument applies more generally to any monotone decision problem and any pair of experiments, as we show in the proof of the following:

Proposition 1. *Let X and Y be two n-dimensional experiments. If Y is conditionally more accurate than X, then for every prior π the optimal expected payoff in experiment Y is greater than or equal to the optimal expected payoff in experiment X. If $n = 1$, then the converse also holds.*

We provide further intuition and discussion on our notion of conditional accuracy in the course of illustrating our second and more important main result, which we are now ready to state.

Welfare Impact of Selection in General Multidimensional Experiments. Using our notion of conditional accuracy, we now identify the crucial property of the family of distributions $(F(\cdot|\theta))_{\theta\in\Theta}$ which guarantees that the correlation structure of the entire vector of selected observations adds or subtracts value to the evaluator's problem as the presample size increases:

Theorem 2. *For a fixed sample size $n \geqslant 1$, an increase in the presample size makes a selected experiment conditionally more (resp. less) accurate if the reverse hazard rate $f(\cdot|\theta)/F(\cdot|\theta)$ is log-supermodular (resp. log-submodular, with support of $f(\cdot|\theta)$ unbounded above for every θ), that is, if for all states θ and $\theta' > \theta$ the reverse hazard rate ratio*

$$\frac{f(\cdot|\theta')/F(\cdot|\theta')}{f(\cdot|\theta)/F(\cdot|\theta)}$$

is increasing (resp. decreasing).

For location experiments, log-supermodularity of the reverse hazard rate is the same as logconcavity of the reverse hazard rate of the basic noise distribution:

Corollary 1. *For a fixed sample size $n \geqslant 1$, an increase in the presample size makes a location experiment conditionally more (resp. less) accurate if the basic noise distribution's reverse hazard rate, $f(\varepsilon)/F(\varepsilon)$, is logconcave (resp. logconvex, with support of f unbounded above) in ε.*

To reconcile this corollary with Theorem 1, observe that the reverse hazard function is the right-sided integral of the reverse hazard rate: $\int_\varepsilon^\infty (f(\varepsilon)/F(\varepsilon))d\varepsilon = -\log F(\varepsilon)$. Thus, the reverse hazard function inherits logconcavity (and logconvexity, if the support of f is unbounded above) of the reverse hazard rate (An, 1998, Lemma 3). Indeed, in all examples of location experiments listed in the previous section, Corollary 1 applies, with the Gumbel distribution again sitting at the boundary between the basic noise distributions for which more selection benefits, and those for which less selection does. In the normal case, the reciprocal of the reverse hazard rate, $F(\varepsilon)/f(\varepsilon) = \int_{-\infty}^{x} e^{\varepsilon^2/2}e^{-t^2/2}dt = \int_{-\infty}^{0} e^{-u^2/2}e^{-u\varepsilon}du$, is logconvex because $e^{-u\varepsilon}$ is logconvex (actually loglinear) and logconvexity is preserved under mixtures (An, 1998, Proposition 3).[28] Thus, the reverse hazard rate is logconcave. In the logistic case, the reverse hazard rate is $f(\varepsilon)/F(\varepsilon) = 1/(e^{\varepsilon}+1)$, which is easily seen to be logconcave. In the Gumbel case, $f(\varepsilon)/F(\varepsilon) = \exp(-\varepsilon)$, a loglinear function. Finally, in the generalized exponential case, $f(\varepsilon)/F(\varepsilon) = (1-e^{-\varepsilon})^{a}e^{-\varepsilon}$, which is easily seen to be logconvex (as $a < -1$).

Let us now illustrate the idea behind Theorem 2, focusing again on simple hypothesis testing. For simplicity, assume sample size $n = 2$. Let X be a selected experiment with presample size $k \geqslant 2$ from a distribution in some family $(F(\cdot|\theta))_{\theta\in\{\theta_L,\theta_H\}}$. In this experiment the evaluator optimally accepts after observing a realization $x \in U$, and rejects otherwise, where the acceptance region U is some upper set in $\mathbb{R}^2_{\geq}$. This is illustrated in Figure 3(a), where we assume that X is a random experiment ($k = 2$) from a normal location family—for every state θ, the distribution $F(\cdot|\theta)$ is normal with mean θ and variance one. The blue curve partitions the domain of X (the unshaded area in the diagram, $\mathbb{R}^2_{\geq}$) into the acceptance region U, the area above the curve, and the rejection region, the area below the curve.[29]

Consider now another selected experiment, Y, with presample size $m \neq k$. Following the same logic used for the one-dimensional case, in order to argue that Y gives a higher payoff than X we must define an acceptance region $V \subseteq \mathbb{R}^2_{\geq}$ for experiment Y, leading to as many false positives and fewer false negatives, that is,

$$\Pr_L(Y \in V) = \Pr_L(X \in U) \quad \text{and} \quad \Pr_H(Y \in V) \geqslant \Pr_H(X \in U). \tag{8}$$

But, unlike the one-dimensional case, where V is uniquely determined by U,[30] there are now many ways to define an upper set V satisfying the equality in (8). Moreover, the inequality in (8) may hold for some choices of V and not for others. How should we define the set V, then?

[28] Balakrishnan, Burkschat, Cramer, and Hofmann (2008) use the same argument to prove, in their Lemma A.2, that the hazard rate $f/(1-F)$ is logconcave in the normal case.

[29] The blue curve is, in fact, a straight line. As is well known, the optimal strategy in a location experiment with i.i.d. normal observations (recall that we are assuming $k = n$ here) only depends on the average observed value, because this is a sufficient statistic for the whole vector of observations.

[30] Recall, going back to our illustration of Theorem 1, that equation (4) uniquely defines the cutoff point $\bar{y}$ from $\bar{x}$.

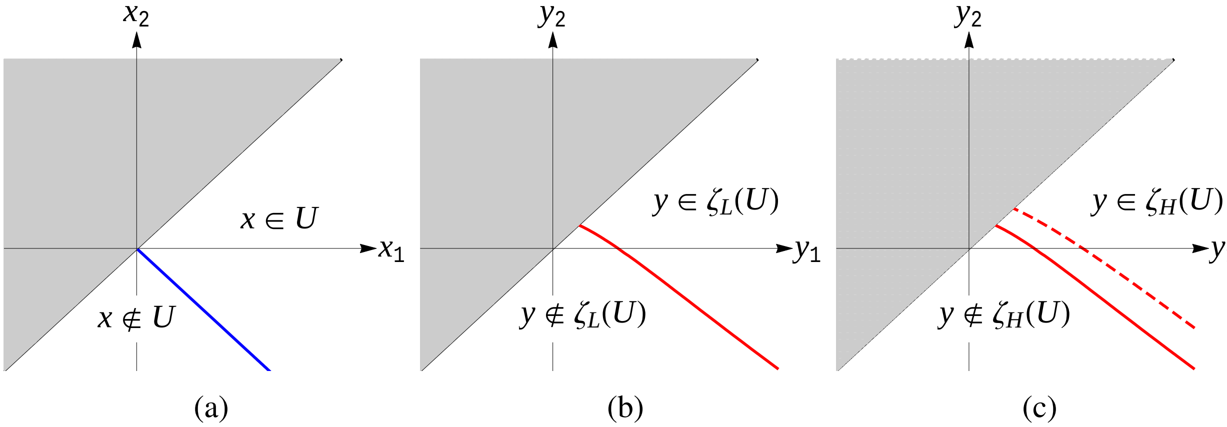

Figure 3: Normal location experiment: increasing presample size increases accuracy.

Our definition of conditional accuracy proves crucial to answering this question. Recalling (3) and (7), for any state θ the function ζ_θ is defined by $\zeta_\theta(x_1,x_2) = (z_1,z_2)$, where

$$F^k(x_1|\theta) = F^m(z_1|\theta) \quad \text{and} \quad \frac{F^{k-1}(x_2|\theta)}{F^{k-1}(x_1|\theta)} = \frac{F^{m-1}(z_2|\theta)}{F^{m-1}(z_1|\theta)}. \tag{9}$$

Letting $V = \zeta_L(U)$, as illustrated in Figure 3(b), the equality in (8) holds by construction. Moreover, since U is an upper set, checking the inequality in (8) simply requires checking that for every realization x we have $z := \zeta_L(x) \leqslant \zeta_H(x)$ or, by (9),

$$\frac{F^m(z_1|\theta_H)}{F^k(x_1|\theta_H)} \leqslant 1 \qquad \text{and} \qquad \frac{F^{m-1}(z_2|\theta_H)/F^{m-1}(z_1|\theta_H)}{F^{k-1}(x_2|\theta_H)/F^{k-1}(x_1|\theta_H)} \leqslant 1.$$

At this point, log-supermodularity and log-submodularity enter the picture. Since both of the above inequalities hold in the limit as x_1 tends to the upper bound of the support of $F^k(\cdot|\theta_L)$ and x_2 tends to its largest possible value (namely x_1), they both hold if their left-hand sides are increasing functions of x_1 and x_2, respectively. As we show in the proof, taking derivatives and simplifying, this simply means that, for each $i = 1,2$,

$$\frac{f(z_i|\theta_H)/F(z_i|\theta_H)}{f(z_i|\theta_L)/F(z_i|\theta_L)} \geqslant \frac{f(x_i|\theta_H)/F(x_i|\theta_H)}{f(x_i|\theta_L)/F(x_i|\theta_L)}, \tag{10}$$

revealing that Y is more accurate than X when either the reverse hazard rate is log-supermodular and $m \geqslant k$ (as then $z_i \geqslant x_i$), or the reverse hazard rate is log-submodular and $m \leqslant k$ (as then $z_i \leqslant x_i$).

As we argued while discussing Corollary 1, the inequality $\zeta_H(\cdot) \geqslant \zeta_L(\cdot)$ holds in the normal case depicted in Figure 3. In panel (c) of that figure, the dashed red curve defines the set $\zeta_H(U)$, the region above the curve. In state θ_H the probability of Y falling above the dashed curve equals the probability that X falls in U. The solid curve, which is the same as in panel (b), lies below the dashed curve, so the probability that Y falls above the solid curve is larger—in other words, the inequality in (8) is satisfied.

4 Applications

Selection bias is an important concern in observational studies, as well as in the practice of controlled experiments. The received statistical and econometric literature on regression and treatment effects is typically concerned with identification issues—finding ways to avoid or at least account for bias. Our analysis offers insights from a complementary angle: for a fixed sample size, does selection provide more or less precise information about the phenomenon of interest? In this section we discuss the implications of our results in three typical applied scenarios.

4.1 Subversion of Randomization in Randomized Controlled Trials

Following Neyman (1923) and Rubin (1974, 1978), consider a population of individuals and two alternative treatments—a default, known treatment 1 and a new treatment 2 whose benefit beyond the default is unknown. Let $X_{t,i}$ denote the potential outcome of individual i when receiving treatment $t \in \{1,2\}$. For simplicity, assume for now that the unknown treatment effect $X_{2,i} - X_{1,i}$ on individual i is the same for every individual i. Thus, the potential outcomes of individual i are

$$X_{1,i} = \varepsilon_i \quad \text{and} \quad X_{2,i} = \theta + \varepsilon_i$$

where the treatment effect θ is only known to belong to a subset $\Theta \subseteq \mathbb{R}$, and ε_i is drawn from a known distribution F with logconcave density f. The evaluator would like to approve the new treatment (action a_2) if the treatment effect exceeds some specific value, and stick with the traditional treatment (action a_1) otherwise.

Enter a researcher, who runs a randomized controlled trial with n treated individuals $i_1, \ldots, i_n$ and as many untreated individuals $i_{n+1}, \ldots, i_{2n}$. The evaluator then observes the following table of experimental results:

Treatment Group	Control Group
$X_{2,i_1} = \theta + \varepsilon_{i_1}$	$X_{1,i_{n+1}} = \varepsilon_{i_{n+1}}$
$\vdots$	$\vdots$
$X_{2,i_n} = \theta + \varepsilon_{i_n}$	$X_{1,i_{2n}} = \varepsilon_{i_{2n}}$

We are interested in comparing two scenarios. In the first scenario, the researcher picks $2n$ individuals at random, and randomly assigns n individuals to each treatment. In this case the control group adds no valuable information, because the distribution of outcomes under the first treatment, namely F, is known. Thus, the experiment boils down to the observation of the treatment group only—in the language used in this paper, the random experiment $X = (X_1, \ldots, X_n)$ where $X_1 = \theta + \varepsilon_{n:n}, \ldots, X_n = \theta + \varepsilon_{1:n}$.

In the second scenario, the researcher has information on the outcome of the first treatment (more generally, on characteristics correlated with this outcome) for $k > 2n$ individuals, and on this basis (i) selects $2n$ individuals for the experiment, and (ii) assigns n individuals to each treatment. Out of the k presampled individuals, the researcher assigns the n individuals with the highest value of X_1 to the treatment group, and the n individuals with the lowest value of X_1 to the control group. Thus, the evaluator observes the following data:

Treatment Group	Control Group
$\theta + \varepsilon_{k:k}$	$\varepsilon_{n:k}$
$\vdots$	$\vdots$
$\theta + \varepsilon_{k-n+1:k}$	$\varepsilon_{1:k}$

How do the two scenarios compare? Clearly, the control group can only add information—and in the second scenario it does, because the random vectors $\varepsilon_{1:k}, \ldots, \varepsilon_{n:k}$ and $\varepsilon_{k-n+1:k}, \ldots, \varepsilon_{k:k}$ are correlated. Thus, the experiment must be at least as accurate as the experiment consisting of the vector of observations in the treatment group—in our language, the random experiment $Y = (Y_1, \ldots, Y_n)$ where $Y_1 = \theta + \varepsilon_{k:k}$, $\ldots$, $Y_n = \theta + \varepsilon_{k-n+1:k}$. By Corollary 1, we can therefore conclude that the evaluator is better off in the second scenario, provided that the reverse hazard rate f/F is logconcave.

4.2 Sample Selectivity in Regression

The estimation of a regression parameter is another prominent example of a monotone decision problem. Here, the set of actions A coincides with the state space Θ, and the evaluator would like to choose an action that is as close as possible to the true state.

Consider a linear regression setup $Y_i = \alpha + \theta X_i + \gamma Z_i + \varepsilon_i$ where X_i is a nonnegative treatment variable, Z_i a vector of covariates, and θ the unknown parameter of interest.[31] We assume that any sample of individuals must be representative of the population, in the sense that the distribution of X and Z in the sample is the same as in the population. More precisely, we assume that X and Z

[31] For simplicity, assume that the other two parameters, α and γ, are known. Our arguments can be easily extended to the case where α is unknown, using Lemma 1 below. The case where the parameter γ is also unknown is more complicated, but we conjecture that our arguments again apply.

define S different strata—so that, across all individuals i in each stratum s, the values of X_i and Z_i are the same—and that any sample of n individuals must contain exactly n_s individuals from each stratum s, with $n_1 + \cdots + n_S = n$.

Our question here is whether a more accurate estimate of θ obtains with random sampling, where the noise terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. draws from some basic noise distribution F, or with the following form of selection: for each stratum s, the n_s individuals in the sample are selected as those with the highest value of ε, among $k_s \geqslant n_s$ randomly drawn individuals from stratum s. Note that we allow stratum-specific presample sizes, reflecting the possibility that selection may be more or less severe depending on the values of X and Z.

Here, too, Corollary 1 provides the answer: the estimate of θ is more accurate in the selected scenario when f/F is logconcave, and less accurate when f/F is logconvex (and the support of f is unbounded above). To see this, consider first each stratum separately. For each individual i in the stratum, the evaluator observes

$$\tilde{Y}_i = \theta + \tilde{\varepsilon}_i,$$

where $\tilde{\varepsilon}_i = \alpha + \gamma Z_i/X_i + \varepsilon_i/X_i$. Since X_i and Z_i are constant within strata, in both the random and the selected case the noise terms $\tilde{\varepsilon}_i$ are drawn from the same basic noise distribution. Moreover, this distribution inherits logconcavity (or logconvexity) of the reverse hazard rate from the basic noise distribution of ε_i, because $\tilde{\varepsilon}_i$ is an affine transformation of ε_i (An, 1998, Corollary 2).

The above argument proves our claim for the case of a single stratum. The result for an arbitrary number S of strata obtains immediately from the fact that, conditional on θ, observations in different strata are independent. Indeed, it is easy to see that combining conditionally more accurate mutually independent experiments results in a conditionally more accurate experiment.[32]

Truncated Regression. It is instructive to contrast our findings with the common applied scenario in which observations on Y, X and Z are only available for individuals whose value of Y exceeds a certain (possibly stratum-specific) threshold—the case of *truncated regression*. In this scenario, for each stratum s the evaluator observes n_s i.i.d. draws from the truncated distribution $\tilde{Y}|\tilde{Y} \geqslant \bar{y}_s$, where $\bar{y}_s$ denotes the left-truncation point for individuals in stratum s. Using the notion of accuracy, and variants of the arguments used to establish Theorem 2, we show (see Theorem 3 in the Supplementary Appendix) that in some important cases (e.g. with a normal basic noise distribution) this type of selection hurts the evaluator, even under the assumption that the truncation points $\bar{y}_s$

[32] More precisely, take a pair of experiments X, X' with, say, sample sizes n and n', and suppose that X is independent of X' in each state. Let Y, Y' be another pair of experiments with sample sizes n and n', again with Y independent of Y' in each state. If Y is conditionally more accurate than X, and Y' conditionally more accurate than X', then the $(n+n')$-dimensional experiment (Y, Y') is conditionally more accurate than (X, X').

are known, while selection based on order statistics is beneficial. This contrast is striking: the two types of selection bear a superficial resemblance, as they both skew upward the data distribution. Through the lens of accuracy, however, they reveal fundamentally different welfare implications. More generally, our notion of conditional accuracy can help in systematically assessing the beneficial or harmful nature of any type of selection—both inside and outside regression contexts.

4.3 Selective Regression Specification

At least since Griliches (1957), economic researchers have sought to address a particular source of bias arising from missing variables. We illustrate here how our analysis can be applied to the situation where the omitted variable has been strategically selected to maximize the bias.[33] Before doing so, we first need to develop a useful add-on to Theorem 1.

Correlated Draws. Our analysis of one-dimensional selected location experiments, which has so far focused on selection among conditionally independent draws, carries over to a relevant case of conditionally correlated draws. In this case, the k presample observations are subject to an identically distributed noise component. More precisely, the k presample observations have the form $\theta + \delta_i + \varepsilon_i$, where the vectors $(\delta_1, \dots, \delta_k)$ and $(\varepsilon_1, \dots, \varepsilon_k)$ are mutually independent, and independent of θ. As before, $\varepsilon_1, \dots, \varepsilon_k$ are i.i.d. draws from a basic noise distribution F with a logconcave density. The additional noise components $\delta_1, \dots, \delta_k$ are identically (but not necessarily independently) distributed with a logconcave density. Note that $\delta_i + \varepsilon_i$ has a logconcave density, as recorded in An (1998, Corollary 1). Selection takes place on the basic noise terms $\varepsilon_1, \dots, \varepsilon_k$. That is, we assume that the evaluator observes $\theta + \delta_{i^*} + \varepsilon_{i^*}$, where $i^* = \arg\max\{\varepsilon_1, \dots, \varepsilon_k\}$, so that the correlation in the draws is introduced after selection.

The key result here is that the added noise components are without consequence for our main question. Whether the evaluator benefits from selection still turns on whether the basic noise distribution's reverse hazard function $-\log F$ is logconcave or logconvex.

Lemma 1. *The selected experiment $\theta + \delta_{i^*} + \varepsilon_{i^*}$ is more (resp. less) accurate than the random experiment $\theta + \delta_i + \varepsilon_i$ if the selected experiment $\theta + \varepsilon_i^*$ is more (resp. less) accurate than the random experiment $\theta + \varepsilon_i$.*

Strategic Variable Omission. Consider a researcher who privately collects a data set (x_i, y_i, z_i) for $i = 1, \dots, n$. Here, x_i is a treatment variable, y_i an outcome variable, and z_i a vector of k covariates.

[33] We provide an explicit game-theoretic foundation for this model in the next section.

The statistical relationship among these variable is captured by the linear model

$$y_i = \theta x_i + z_{1i} + \cdots + z_{ki} + v_i, \tag{11}$$

where v_i is a random noise term. We assume that the marginal effects of covariates are all known, say because the covariates are familiar. Covariates have been scaled so these effects are identical.

An evaluator is concerned that treatment need not be random. The evaluator correctly understands that a regression of z_j on x yields the population relationship

$$z_{ji} = \varepsilon_j x_i + u_{ji}, \tag{12}$$

where u_{ji} is noise. However, the evaluator does not know the realized regression coefficients $\varepsilon_1, \ldots, \varepsilon_k$.

Assume that $\theta, \varepsilon, u, v, x$ are independent stochastic variables, satisfying $E\left[u_{ji}\right] = E\left[v_i\right] = 0$. Moreover, $\varepsilon_1, \ldots, \varepsilon_k$ are i.i.d. draws from a distribution F admitting a logconcave density. Also, the vectors $u_1, \ldots u_k$ are identically distributed. Furthermore, assume that one of the following sets of assumptions applies:

1. The treatment variable x is known and non-stochastic. Noise terms u_{ji} and v_i have logconcave densities.
2. The treatment variable x is stochastic, unknown to the evaluator. The term $\left(x^T x\right)^{-1} x^T \left(u_j + v\right)$ has a logconcave density.

One of the control variables is omitted when the researcher uses the realizations to compute an estimate of θ. We compare the case where this omission is of a random control variable to the case where the omitted variable is selected to maximize the estimate. Selection is performed when the researcher knows $\varepsilon_1, \ldots, \varepsilon_k$, but before obtaining the actual data set (x, y, z) for the estimation of θ. The bias thus results from the researcher's prior analysis of covariates, or from a non-random assignment procedure.

Proposition 2. *Omission of variable j results in estimate $\theta + \varepsilon_j + \delta_j$, where*

$$\delta_j = \left(x^T x\right)^{-1} x^T \left(u_j + v\right).$$

This proposition reduces the problem to our earlier analysis. Our technical assumptions mean that we can apply Lemma 1 in order to characterize whether selected omission benefits or harms the evaluator.[34] The answer turns on whether $-\log F$ is logconcave or logconvex.

[34] We have directly assumed that ε_j are i.i.d. with a logconcave density, independent of vector δ. We have also imposed two alternative assumptions to guarantee that the terms δ_j are identically distributed with logconcave density.

5 Strategic Selection

Sample selection of the sort considered above naturally arises as an equilibrium phenomenon in a strategic setting where the experiment is carried out by a researcher who desires the evaluator to take higher actions. Taking the researcher's payoff into account allows us to discuss the non-trivial impact on the researcher's own welfare when sample selection becomes easier. Initially, we exogenously fix both size n of the reported sample and size k of the researcher's presample. Subsequently, we endogenize these features, while also imposing more restrictive assumptions on signals and payoffs in order to facilitate the analysis.

5.1 Selective Sampling Game

In many settings, size n of the reported sample may be determined by forces outside the model. For instance, costly or limited attention from the evaluator may cap the number of items that the evaluator is willing to inspect. In some areas, a sample size may be guided by standards unrelated to the specific problem. Similarly, natural constraints may endow the researcher with a presample of fixed and known size $k \geqslant n$.

Consider the following timeline:

Stage 1. Researcher privately observes presample $(x_1,\ldots,x_k)$ and then chooses a subset $\mathscr{I} \subseteq \{1,\ldots,k\}$ of size n.

Stage 2. The evaluator observes the vector $(x_i)_{i\in\mathscr{I}}$ and then chooses an action in A.

Payoffs to the evaluator are as specified before. The researcher's payoff is $v(a)$ where v is a strictly increasing function from the ordered set A to the reals.

Proposition 3. *There exists a Bayes Nash equilibrium where the researcher always selects the n greatest realizations from presample $(x_1,\ldots,x_k)$. The evaluator's map from data to actions is as analyzed before.*

We use the Bayes Nash equilibrium concept since the researcher has private information. On the other hand, with the proposed strategy, no sample $(x_1,\ldots,x_n)$ is outside the support, so there is no reason to discuss any refinement of off-path beliefs.

Note that the evaluator's strategy is precisely the one we have analyzed until now. The researcher's strategy is a best response because it generates the highest possible realization of sample vector $(x_1, \ldots, x_n)$, and the evaluator then responds with no smaller action. It is technically possible to construct examples of equilibria in which either the evaluator adopts a non-monotonic strategy or the researcher does not always report the n highest of the k observations. These other equilibria are more intricate, and perhaps less natural in light of the researcher's goal to maximize the evaluator's response, so we choose to disregard them for now. When the evaluator is uncertain about the extent of selection k, there are situations in which maximal selection is no longer an equilibrium—see the discussion at the end of Section 6.1; Proposition 7 in the Supplementary Appendix analyzes equilibria without maximal selection.

If the researcher could somehow ex post choose to reveal the entire presample, there are cases when the researcher would wish to do so. Generally, these cases arise when the hidden part of the sample has a favorable realization, beating the posterior odds assumed by the evaluator. This unraveling effect is well known in the literature on strategic disclosure, at least since Grossman (1981) and Milgrom (1981). Our assumption is that the evaluator can ex ante commit to study precisely n observations before making a decision, preventing such unraveling. We will endogenize n in the next subsection.

The result is robust to a natural modification of the game for the case of location problems, where $x_i = \theta + \varepsilon_i$. We can modify the first stage of the game by allowing the researcher to observe the noise terms $(\varepsilon_1, \ldots, \varepsilon_k)$ rather than the outcomes. Of course, for any realization of the state, maximal selection in $(\varepsilon_1, \ldots, \varepsilon_k)$ is equivalent to maximal selection in $(x_1, \ldots, x_k)$. The result is therefore unchanged.[35]

Equilibrium Impact of Selection on Researcher's Welfare. Holding fixed the sample size n, does more selection improve the researcher's situation? If the evaluator were unaware of an increase in presample size k, a direct effect would benefit the researcher. Increasing k would directly raise the maximally selected sample in the sense of first-order stochastic dominance. Holding fixed the evaluator's response, higher actions would result, to the benefit of the researcher.

We are more interested in the equilibrium effect. When both the evaluator and the researcher know that selection is easier—that is, k is larger—is the researcher necessarily better off? Does this depend on whether the evaluator is better off?

We can immediately note that the researcher is entirely indifferent to k in the special case of

[35] For example, this setup can be interpreted as subversion of randomization in randomized controlled trials, with known effect from the default treatment.

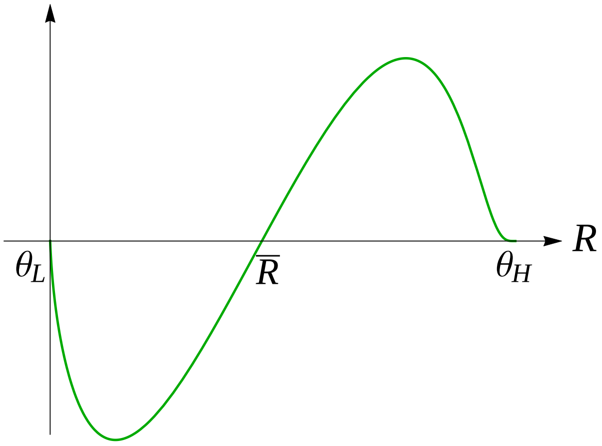

Figure 4: Researcher gain from selection, as evaluator's preference varies.

a location experiment with Gumbel basic noise distribution. As we know, in this case all selected experiments (with the same sample size) are equally conditionally accurate. By definition of conditional accuracy, this means that any joint distribution on $\Theta \times A$ that the evaluator can induce with a strategy in a selected experiment, the evaluator can also induce in another selected experiment. Hence, in particular, the distribution over A induced by the evaluator's optimal strategy is independent of the presample size k.

For a more interesting example, consider a simple hypothesis testing problem and a one-dimensional location experiment with a normal basic noise distribution. For convenience, from now on normalize the evaluator's payoffs, so that the payoff from acceptance in each state θ is simply equal to θ, and rejection leads to a safety payoff R. Thus,

$$\theta_L = u(\theta_L, a_2) \leqslant \underbrace{u(\theta_L, a_1) = u(\theta_H, a_1)}_{=R} \leqslant u(\theta_H, a_2) = \theta_H.$$

We investigate the comparative statics of the researcher's welfare gain from selection, as the evaluator's safety payoff R varies between θ_L and θ_H. Letting p denote the prior probability of state θ_H, the researcher's payoff is the ex ante probability of acceptance, that is,

$$p[1 - F(\bar{x} - \theta_H)] + (1 - p)[1 - F(\bar{x} - \theta_L)],$$

where $\bar{x}$ is the cutoff point chosen by the evaluator.

Figure 4 illustrates the comparative statics result when $n = 1$ and $k = 2$, and the prior $p = 1/2$. The green curve represents the researcher's payoff difference between the selected and random experiment. When R is low, the researcher suffers from selection. When R is high, the researcher benefits.[36] Intuitively, when R is low, the evaluator is more willing to accept. Providing more information (with higher k in the normal example), allows the evaluator to reject more often (in state θ_L). This is not in the researcher's interest. Conversely when R is high.[37]

5.2 Data Production Game

In the previous section we exogenously fixed sample size n and presample size $k \geqslant n$. Here, we endogenize the choice of n and k. Specifically, we consider two stages preceding the selection game studied before:

Stage -1. Evaluator chooses the sample size n and decides whether a presample is permitted.

Stage 0. Researcher privately chooses presample size $k \geqslant n$, or opts out of the game. If a presample is not permitted, the researcher only considers $k = n$ or opting out.

The researcher's gross payoff is still $v(a)$, but the researcher must bear cost $C(k)$ for obtaining the presample. We assume that this cost function is increasing and convex—if we restrict attention to natural numbers k, convexity means that $C(k+1) - C(k)$ is increasing in k. Opting out is free, so $C(0) = 0$. The choice of k is private, and we assume that there is no credible way to directly signal any information about it. In equilibrium, the evaluator correctly anticipates k. We focus on pure-strategy equilibria, consistent with our statistical analysis so far where k was fixed.[38]

Payoffs to the evaluator are as specified before. We endow the evaluator with an option to completely forbid sample selection, assuming that such a rule can be perfectly enforced. This assumption allows the evaluator who stands to suffer from sample selection to completely rule it out. More interestingly, as we will show, the evaluator will in some circumstances prefer to tolerate sample selection, despite the availability of such a strong instrument. For simplicity, we ignore the evaluator's cost of studying larger samples. In case the researcher opts out, we assume that the evaluator must rely on an expensive alternative source of information—we will use this only to break ties in stages -1 and 0.

[36] At $R = (\theta_L + \theta_H)/2$, the researcher benefits from selection. Without selection, in this symmetric case, the evaluator chooses the cutoff $\bar{x}$ inducing as many false positives as false negatives: $1 - F(\bar{x} - \theta_L) = F(\bar{x} - \theta_H)$. Maximal selection introduces an asymmetry in the distribution that leads the evaluator to optimally choose $\bar{x}$ such that $1 - F(\bar{x} - \theta_L) > F(\bar{x} - \theta_H)$, that is, more false positives are tolerated.

[37] This intuition is close to that in Johnson and Myatt (2006).

[38] We later discuss the implications of uncertainty regarding k.

Following stages -1 and 0 above, we assume sequential rationality: the researcher selects the highest observations, and the evaluator best-responds to that selection, given the conjectured presample size, as in Proposition 3. Since the presample size k is privately chosen in stage 0, we need an expression for the researcher's expected payoff after deviating. With sample size n, actual presample size k, and $\hat{k}$ the presample size conjectured by the evaluator, we let $V\left(n,k,\hat{k}\right)$ denote this expected payoff. The distribution over observations here is governed by k and maximal selection by the researcher, but the evaluator best-responds to n out of $\hat{k}$.

Equilibrium Choice of Presample Size. Both parties observe the evaluator's choice of n in stage -1. We now consider stage 0 in the nontrivial case where greater presamples $k > n$ are permitted.

We restrict attention to simple hypothesis testing, and we also focus this first analysis on the case $n = 1$. The evaluator accepts when the realization x exceeds cutoff point $\bar{x}$, an increasing function of the conjectured presample size $\hat{k}$. In stage 0, the researcher chooses k in order to maximize the payoff $V\left(n,k,\hat{k}\right) - C(k)$, that is,

$$p\left[1 - F^k\left(\bar{x} - \theta_H\right)\right] + (1-p)\left[1 - F^k\left(\bar{x} - \theta_L\right)\right] - C(k). \tag{13}$$

Considering a deviation from equilibrium, the researcher has a potential gain through the upward shift of the realized observation x. This is to be weighed against the cost of looking at more subjects, when already looking at k.

Proposition 4. *The researcher's objective function is concave in k. The first order condition (assuming $k \geqslant 1$ is a real number) is given by*

$$-p\log\left(F\left(\bar{x} - \theta_H\right)\right)F^k\left(\bar{x} - \theta_H\right) - (1-p)\log\left(F\left(\bar{x} - \theta_L\right)\right)F^k\left(\bar{x} - \theta_L\right) = C'(k). \tag{14}$$

The researcher opts out if this k provides lower payoff in (13) than the probability of acceptance when the evaluator employs the alternative information source.

The equilibrium described in Proposition 4 exhibits a rat race effect: when the evaluator correctly anticipates a higher degree k of selection, the researcher's cost $C(k)$ to manipulate the experiment is partly wasted. To see the cleanest instance of this, consider the Gumbel example. The private choice of k tempts the researcher to choose $k > 1$ (we prevent opt-out by assuming that the evaluator is significantly less likely to accept without the researcher's sample). The evaluator's acceptance probability is independent of $k \geqslant 1$. The researcher's total payoff, having chosen $k > 1$, is then less than if the researcher could be tied to the mast, unable to augment k. Beyond the Gumbel example, costs of manipulation could even harm a researcher who already lost welfare due to sample selection.

The researcher's best response k may increase or decrease with the evaluator's cutoff $\bar{x}$, depending on parameters. The sign of this slope depends on the sign of the derivative of the left hand side in (14) with respect to $\bar{x}$,

$$-p\left(1+k\log\left(F\left(\bar{x}-\theta_H\right)\right)\right)F^{k-1}\left(\bar{x}-\theta_H\right)f\left(\bar{x}-\theta_H\right) \\ -\left(1-p\right)\left(1+k\log\left(F\left(\bar{x}-\theta_L\right)\right)\right)F^{k-1}\left(\bar{x}-\theta_L\right)f\left(\bar{x}-\theta_L\right),$$

which is positive when $\bar{x}$ is sufficiently small, as happens when the safety payoff R is small. In that case, the best response k is an increasing function of $\bar{x}$. Conversely, when R is large, the best response k is a decreasing function of $\bar{x}$.[39]

This comparative statics result allows us to discuss the evaluator's optimal commitment to an ex-post suboptimal acceptance standard. If R is large, the researcher's best response is downward sloping. If the evaluator stands to gain welfare from selection (e.g., if F has logconcave reverse hazard function), it is optimal for the evaluator to commit to reduce the acceptance standard below the Nash level in order to induce the researcher to increase k. Conversely, when R is small.

Equilibrium Choice of Sample Size. We finally turn to stage -1. If the researcher ends up considering a presample of size k, why should the evaluator not request to see all the evidence? The evaluator could increase n until it hits k, or the evaluator could forbid presampling. The argument against this is that a freedom to presample can create value for the researcher, and hence make the researcher more willing to produce the sample. We will show by a simple example that this can be consistent with equilibrium. The example serves as proof of the more general concept.

For this example, we consider again simple hypothesis testing with normal noise. To simplify the analysis, we consider parameters that satisfy *equipoise*: ex ante, the evaluator is indifferent among acceptance and rejection.[40] Slightly stronger, we assume $p = 1/2$ and $R = \left(\theta_L + \theta_H\right)/2$.

Suppose that $\hat{k} = n$, that is, the evaluator anticipates no selection. Then the evaluator accepts if and only if $(1/n)\sum_{i=1}^{n} x_i \geqslant \left(\theta_L + \theta_H\right)/2$. By symmetry of this setting, the equilibrium probability of acceptance is $1/2$, independent of n. The researcher's equilibrium payoff (if $k = n$) from the evaluator's action is then invariant to k, that is, $V\left(n,n,n\right) = 1/2$. However, participation cost $C\left(k\right)$ rises with $k = n$. In this example, we make the natural assumption that the evaluator accepts with probability $1/2$ if the researcher does not produce evidence. By implication, the researcher opts

[39] It can be easily verified that the best reply of the researcher is increasing for $\bar{x} < \theta_L + F^{-1}(e^{-1/k})$ and decreasing for $\bar{x} > \theta_H + F^{-1}(e^{-1/k})$.

[40] The condition of equipoise, requiring experimental subjects to be indifferent between treatment and control, is an ethical prerequisite that is often required for carrying out a randomized experiment. See Freedman (1987).

out. Were the researcher still to participate, the evaluator's welfare would rise with $k = n$, as more evidence allows acceptance to grow more likely in state H and less likely in state L (at equal rates).

Consider now the case $n = 1$ and $k = 2$. As noticed in footnote 36, the evaluator accepts with a chance strictly greater than $1/2$, to the interest of the researcher, so $V(1,2,2) > 1/2$. Thus, the researcher is willing to participate when $n = 1$ and presampling is allowed, provided that $C(2) < V(1,2,2) - 1/2$. Recall from Proposition 4 that the researcher's objective function is concave in k, so for $C(3)$ sufficiently large, the researcher will not deviate to choose $k = 3$ when $\hat{k} = 2$ is conjectured. Likewise, for $C(3)$ sufficiently large, the researcher will opt out if $n = 2$ or larger. Finally, when alternative information is expensive, the evaluator will prefer the case $n = 1, k = 2$ over a researcher who opts out.

The example discussed above uncovers a mechanism whereby the evaluator benefits from fixing the sample size and tolerating sample selection from a presample larger than the sample—thus *committing* not to look at more data. Intuitively, the bias (in terms of increased acceptance) that results from selective reporting benefits the researcher, relaxes the individual rationality constraint, and thus induces the researcher to observe a presample larger than the required sample. In turn, the evaluator also benefits in our leading case with logconcave reverse hazard rate. In other words, selective disclosure arises endogenously in this model—the evaluator benefits from blocking unraveling. Both parties are better off with strategic sample selection, even when the evaluator has the possibility to forbid it.

6 Extensions

So far, we considered situations in which the evaluator perfectly predicts the extent of selection, for example because the parameters of the model—such as the researcher's bias and cost of presampling—are known. This is the most optimistic scenario when evaluating the impact of selection. In this section we relax these assumptions and consider more realistic scenarios. We sketch extensions of the analysis to situations where the evaluator may be uncertain about the correct extent of selection k, or even fail altogether to anticipate any selection.

6.1 Uncertain Selection

Sometimes, the extent of selection is known, for instance because a study has been conducted in n of k possible groups, or because a student has solved n of k given problems. Nevertheless, in other settings it seems natural that the evaluator is uncertain about k, the extent of selection. Seeing a

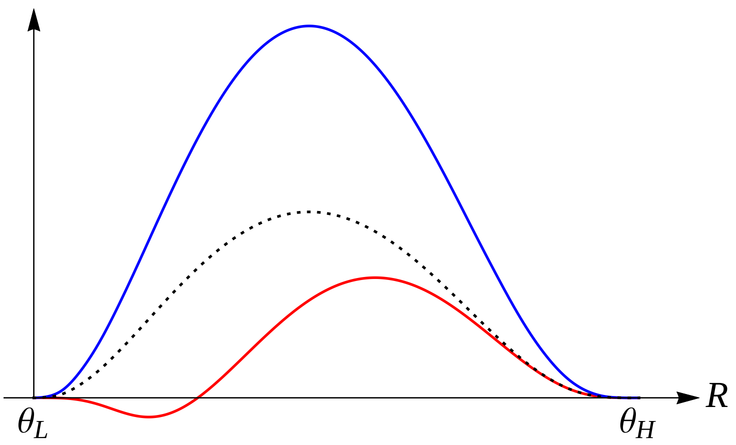

Figure 5: Evaluator's gain over certain $k = 1$ in normal hypothesis testing, as a function of the safety payoff R. Blue curve: certain selection $k = 2$. Red curve: equal chance of $k = 1$ and $k = 2$.

sample of size n, the presample size $k \geqslant n$ appears random. For instance, uncertainty arises with endogenous sample selection when the evaluator does not know precisely the costs and preferences of the researcher.

As one should expect, uncertainty about the extent of selection tends to harm the evaluator. This is particularly evident in the Gumbel case, where known selection leaves the evaluator indifferent:

Proposition 5. *Consider a location experiment with sample size $n = 1$, and suppose that the basic noise distribution is Gumbel. (i) If k follows a logconcave Gamma distribution on $(1, \infty)$, then the noise distribution in the selected experiment has a logconcave density. (ii) For any non-degenerate distribution of k, when the evaluator sees the maximally selected outcome, the evaluator's welfare is strictly lower than when k is known.*

Extending beyond the Gumbel case, the lack of information about the extent of selection k is a force that tends to reduce the evaluator's welfare. While this is an important caveat to our earlier findings where the evaluator sometimes benefits from selection, our main results are only partly overturned by uncertainty. Figure 5, for instance, compares the welfare impact of certain selection ($k = 2$) and uncertain selection (equal chance of $k = 1$ and $k = 2$) in simple hypothesis testing with a normal basic noise distribution. As we know from Theorem 1, certain selection benefits the evaluator for all parameter values, as shown by the blue curve. Now suppose that nature decides whether $k = 1$ or $k = 2$. If the evaluator could observe nature's move, the gain over a random experiment would be measured by the dotted curve, which is exactly half the blue curve (since there are equal chances of $k = 1$ and $k = 2$). Under uncertainty, the evaluator must fare worse—the red curve lies below the dotted curve. Still, the evaluator often fares better than in a random experiment: this occurs in the realistic case where the evaluator a priori does not favor acceptance too strongly (R is not too small). In fact, if the evaluator a priori strongly favors rejection, uncertainty has almost no impact—the red and dotted curves almost coincide for R sufficiently large.

Our earlier characterizations also remain robust to the introduction of a *small* amount of uncertainty. Indeed, the evaluator can behave as if k is known. If the distribution over random k converges to this degenerate assumption, the evaluator's payoff converges to the payoff where it really is known, by continuity of expected payoffs. Responding optimally to uncertainty can only improve the evaluator's payoff.

Returning to Proposition 5, part (i) of the result is less trivial than it may appear. The mixture (over k) of logconcave distributions generally need not be logconcave.[41] Once logconcavity fails, then we cannot guarantee existence of the monotone equilibrium from Proposition 3. There will now exist signals $x' > x$ such that, assuming maximal selection, the optimal action after x' is lower than the action after x. A strategic researcher who has both x and x' in the presample will then be tempted not to reveal the maximal signal. We illustrate the construction of a Bayes Nash equilibrium when the MLR property fails, and strategic selection is not everywhere maximal, in the Supplementary Appendix.

6.2 Unanticipated Selection

Consider an unwary evaluator who wrongly anticipates a smaller presample size than true. Holding fixed the true sample size, clearly the evaluator is generally worse off by being unwary than being rational. More interestingly, it is ambiguous whether an unwary evaluator gains or loses when the true sample size is larger than expected. If a rational evaluator would benefit from selection, this gain might be greater than the cost of irrationality.

In an important benchmark case, we find that the unwary evaluator is exactly indifferent to an increase of selection from $k = 1$ to $k = 2$. Consider again a situation of equipoise, whereby at the prior the evaluator is indifferent between accepting and rejecting. Suppose that the basic noise distribution F is symmetric, so that for some ε_0 we have $F(\varepsilon_0 + \varepsilon) = 1 - F(\varepsilon_0 - \varepsilon)$ for all ε. Start from the acceptance cutoff that is optimal in the random experiment, namely $\bar{x} = \varepsilon_0 + (\theta_L + \theta_H)/2$, and consider how selection with $k = 2$ affects an unwary evaluator who maintains the acceptance standard unchanged at $\bar{x}$. The probability of acceptance clearly increases, in both states. The

[41] Our footnote to the proof of Proposition 5 notes the existence of such examples.

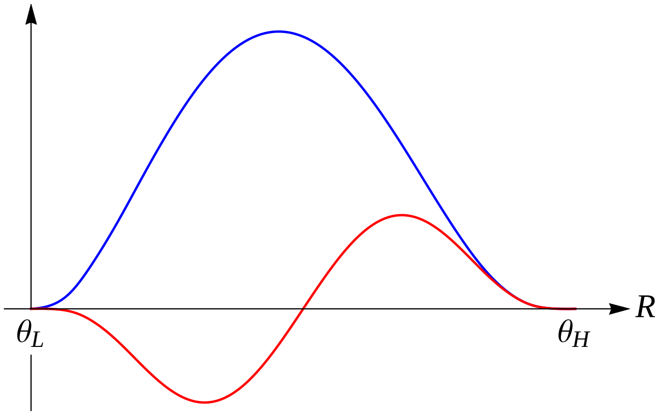

Figure 6: Evaluator's gain from anticipated (blue) and unanticipated (red) selection, compared to a random experiment.

resulting change in the evaluator's payoff equals

$$-(1-p)\underbrace{\left[F(\bar{x}-\theta_L)-F^2(\bar{x}-\theta_L)\right]}_{\text{increase in false positives}}(R-\theta_L)+p\underbrace{\left[F(\bar{x}-\theta_H)-F^2(\bar{x}-\theta_H)\right]}_{\text{reduction in false negatives}}(\theta_H-R).$$

By equipoise, $(1-p)(R-\theta_L)=p(\theta_H-R)$. Thus, false positives and false negatives are equally costly for the evaluator. By symmetry, $F(\bar{x}-\theta_L)+F(\bar{x}-\theta_H)=1$. Thus, the increase in false positives exactly equals the reduction in false negatives. We conclude that the unwary evaluator, who anticipates no selection ($k=1$), is indifferent between no selection and selection with $k=2$.

Can the unwary evaluator *strictly benefit* from selection? Figure 6 shows that this can indeed be the case, in the important special case of a normal location experiment and in the realistic scenario where the evaluator a priori favors rejection—with parameters $p=1/2$ and $R>(\theta_L+\theta_H)/2$. As we know from Theorem 1, a rational evaluator benefits from selection—this is illustrated by the blue curve in the figure, where we assume $k=2$. But, as the red curve shows, an unwary evaluator who would reject at the prior, and wrongly anticipates random data, $k=1$, also benefits from observing selected data with $k=2$. Intuitively, when the upper tail of the basic noise distribution is sufficiently thin—as in the normal case—and R is large enough, the cutoff point $\bar{x}$ is relatively near the upper bound of the (either random or selected) data distribution in the low state. This makes the increase in false positives relatively small. The cutoff point $\bar{x}$ is relatively farther away from the upper bound of the data distribution in the high state, because this distribution stochastically dominates the data distribution in the low state. This makes the decrease in false negatives—the vertical distance between F and F^2 at $\bar{x}$ in the high state—relatively larger. More generally, for every $k \geqslant 2$ there exists a critical threshold on R above which an unwary evaluator benefits from selection with presample size k.

7 Conclusion

Contrary to naive intuition, sample selection does not necessarily damage the evaluator. Increased selection benefits when the data distribution features a log-supermodular reverse hazard rate. For location experiments, this property corresponds to logconcavity of the noise distribution's reverse hazard rate—with a top tail thinner than the extreme value Gumbel distribution and the bottom tail thicker than Gumbel—a condition satisfied by normal noise. Sample selection is detrimental in the log-submodular case—for location experiments, a logconvex reverse hazard rate of the noise distribution. Adding uncertainty or unawareness of selection adds further damage. We also characterize situations in which the researcher ends up suffering from information manipulation like in a rat race, even if we abstract away from the cost of acquiring information. At the same time, we exhibit cases in which the evaluator willingly chooses to allow sample selection in order to incentivize the researcher to provide more evidence. We also develop a generally applicable methodology for comparing the value of information in multidimensional experiments with correlated observations.

The Supplementary Appendix discusses a number of extensions. First, we analyze the impact of other forms of selection such as truncation. Second, drawing on extreme value theory, we analyze the case of extreme selection, with k tending to infinity. We show that for a large class of noise distributions the evaluator achieves full information—in simple hypothesis testing, zero false positives and false negatives—in the limit. Third, we develop the equilibrium analysis under uncertain selection, allowing for non-monotone strategies. Fourth, we sketch how to generalize our analysis to encompass comparisons of experiments that are not ranked by accuracy or conditional accuracy. Finally, we suggest an approach for developing practical criteria to assess the impact of selection in empirical data.

We leave to future work the design of experiments and policy responses in the presence of strategic selection. A natural starting point in this direction is Chassang, Padró i Miquel, and Snowberg's (2012) characterization of experimental design when outcomes are affected by experimental subjects' unobserved actions. Also, given the work by Allcott (2015) on site selection bias, another open question is a general characterization of the impact of selection challenging external validity in the presence of heterogeneous treatment effects.

A Proofs

Proof of Theorem 1. Let $\phi(\varepsilon) = -\log(-\log(F(\varepsilon)))$ and observe that for all ε and $k \geqslant 1$ the horizontal distance between F^k and F, namely $(F^k)^{-1}(F(\varepsilon)) - \varepsilon$, is the same as the horizontal distance between the double-log transformations of F^k and F, namely $\varphi^{-1}(\varphi(\varepsilon) + \log(k)) - \varepsilon$. The derivative of the latter distance is

$$\frac{\varphi'(\varepsilon)}{\varphi'(\varphi^{-1}(\varphi(\varepsilon)+\log(k)))} - 1, \tag{15}$$

which is negative (resp. positive) for all ε and $k \geqslant 1$ if and only if φ is convex (resp. concave). By Theorem 5.2 in Lehmann (1988), F^k is more accurate than F if and only if φ is convex. □

Proof of Proposition 1. The second statement was proved by Lehmann (1988, Theorem 5.1). We now prove the first statement. Since the evaluator's optimal strategy is monotone, the evaluator partitions D into a sequence of sets $(D_1, \dots, D_L)$ such that, for every ℓ, the set $U_\ell = D_\ell \cup \cdots D_L$ is an upper set in D, and chooses action a_ℓ if and only if the observed realization of X belongs to D_ℓ. For every $\ell = 1, \dots, L$ define $U_\ell = D_\ell \cup \cdots \cup D_L$. Then the optimal expected payoff, $\int_\Theta \sum_\ell \Pr_\theta(X \in D_\ell) u(a_\ell, \theta)\pi(\theta) d\theta$, can be rewritten, summing by parts and disregarding constants, as

$$\int_\Theta \sum_{\ell<L} \Pr_\theta(X \in U_{\ell+1}) \big[u(\theta, a_{\ell+1}) - u(\theta, a_\ell)\big] \pi(\theta) d\theta.$$

In order to prove the lemma it therefore suffices to show that if Y is conditionally more accurate than X then we can exhibit nested upper sets $V_2 \supseteq \cdots \supseteq V_L$ such that, for every $\ell = 1, \dots, L-1$ and every state θ, the difference

$$\Pr_\theta(Y \in V_{\ell+1}) - \Pr_\theta(X \in U_{\ell+1}) \tag{16}$$

is nonpositive if $\theta \leqslant \theta_\ell$ and nonnegative if $\theta > \theta_\ell$. Define $V_{\ell+1} = \zeta_{\theta_\ell}(U_{\ell+1})$ for every $\ell = 1, \dots, L-1$. Then we can rewrite the difference in (16) as

$$\Pr_\theta(Y \in \zeta_{\theta_\ell}(U_{\ell+1})) - \Pr_\theta(Y \in \zeta_\theta(U_{\ell+1})).$$

For $\theta \leqslant \theta_\ell$ the difference is nonpositive, because $\zeta_\theta(\cdot) \leqslant \zeta_{\theta_\ell}(\cdot)$ in this case. For $\theta > \theta_\ell$ it is nonnegative, because then $\zeta_\theta(\cdot) \geqslant \zeta_{\theta_\ell}(\cdot)$. □

Proof of Theorem 2. For each state θ, let ζ_θ be the function such that $\zeta_\theta(X)$ has the same distribution as Y. Thus, $\zeta_\theta(x_1, \dots, x_n) = (z_1, \dots, z_n)$, where $z_1, \dots, z_n$ are defined by

$$F^k(x_1|\theta) = F^m(z_1|\theta) \quad \text{and} \quad \frac{F^{k-i+1}(x_i|\theta)}{F^{k-i+1}(x_{i-1}|\theta)} = \frac{F^{m-i+1}(z_i|\theta)}{F^{m-i+1}(z_{i-1}|\theta)} \quad \text{for } i = 2, \dots, n.$$

Before proceeding with the proof, we make the following preliminary observation: since the densities of the distributions $F^k(\cdot|\theta)$ and $F^m(\cdot|\theta)$ have the same support, namely the support of $F(\cdot|\theta)$, as x_1 converges to the upper bound of this support, so does z_1. Similarly, for every $i = 2, \dots, n$ and every x_{i-1}, as x_i converges to its largest possible value, namely x_{i-1}, z_i converges to z_{i-1}.

Fix two states θ' and $\theta'' > \theta'$, and let $z' = \zeta_{\theta'}(x)$ and $z'' = \zeta_{\theta''}(x)$ for brevity. We must prove that under either condition in the theorem ($m \geqslant k$ and the reverse hazard rate is log-supermodular, or $m \leqslant k$ and the reverse hazard rate is log-submodular) for every x we have $z'' \geqq z'$, or equivalently

$$F^m(z_1'|\theta'') \leqslant F^m(z_1''|\theta'') \quad \text{and} \quad \frac{F^{m-i+1}(z_i'|\theta'')}{F^{m-i+1}(z_{i-1}''|\theta'')} \leqslant \frac{F^{m-i+1}(z_i''|\theta'')}{F^{m-i+1}(z_{i-1}''|\theta'')} \quad \text{for } i = 2, \dots, n.$$

Plugging the definition of z'', we can rewrite these inequalities as

$$F^m(z_1'|\theta'') \leqslant F^k(x_1|\theta'') \quad \text{and} \quad \frac{F^{m-i+1}(z_i'|\theta'')}{F^{m-i+1}(z_{i-1}''|\theta'')} \leqslant \frac{F^{k-i+1}(x_i|\theta'')}{F^{k-i+1}(x_{i-1}|\theta'')} \quad \text{for } i = 2, \dots, n.$$

But, for every $i = 2, \dots, n$, if $(z_1'', \dots, z_{i-1}'') \geqq (z_1', \dots, z_{i-1}')$ then the denominator of the left-hand side of the second inequality becomes smaller, and hence the left-hand side of the inequality larger, if we replace z_{i-1}'' with z_{i-1}'. Rearranging terms, we conclude that it suffices to prove that

$$\frac{F^m(z_1'|\theta'')}{F^k(x_1|\theta'')} \leqslant 1 \quad \text{and} \quad \frac{F^{m-i+1}(z_i'|\theta'')}{F^{k-i+1}(x_i|\theta'')} \leqslant \frac{F^{m-i+1}(z_{i-1}'|\theta'')}{F^{k-i+1}(x_{i-1}|\theta'')} \quad \text{for } i = 2, \dots, n. \tag{17}$$

By the preliminary observation, as x_1 tends to the upper bound of the support of the density associated to $F(\cdot|\theta')$, so does z_1. Thus, under either condition in the theorem ($m \geqslant k$, or $m \leqslant k$ and the support of $F(\cdot|\theta')$ is unbounded above), the left-hand side of the first inequality in (17) tends to a number no greater than one. This implies that the first inequality in (17) holds if the left-hand side of the inequality increases with x_1, that is, differentiating with respect to x_1 and dropping the positive denominator in the derivative,

$$mF^{m-1}(z_1'|\theta'')f(z_1'|\theta'')\frac{dz_1'}{dx_1}F^k(x_1|\theta'') \geqslant kF^{k-1}(x_1|\theta'')f(x_1|\theta'')F^m(z_1'|\theta''). \tag{18}$$

But, by definition of z',

$$\frac{dz_1'}{dx_1} = \frac{kF^{k-1}(x_1|\theta')f(x_1|\theta')}{mF^{m-1}(z_1'|\theta')f(z_1'|\theta')}.$$

Plugging the latter in (18) and simplifying, we conclude that the first inequality in (17) holds if

$$\frac{f(z_1'|\theta'')/F(z_1'|\theta'')}{f(z_1'|\theta')/F(z_1'|\theta')} \geqslant \frac{f(x_1|\theta'')/F(x_1|\theta'')}{f(x_1|\theta')/F(x_1|\theta')},$$

which in turn follows from log-supermodularity (resp. log-submodularity) of the reverse hazard rate when $m \geqslant k$ (resp. $m \leqslant k$), because $m \geqslant k$ implies $z_1' \geqslant x_1$ (resp. $m \leqslant k$ implies $z_1' \leqslant x_1$).

Again by the preliminary observation, for every $i = 2, \dots, n$ and every x_{i-1}, as x_i converges to x_{i-1}, z_i' converges to z_{i-1}'. Thus, as before, under either condition in the theorem the left-hand side of the second inequality in (17) tends to a number no greater than the right-hand side. The second inequality in (17) then holds if its left-hand side increases with x_i. Differentiating with respect to x_i and simplifying, as before, we obtain

$$\frac{f(z_i'|\theta'')/F(z_i'|\theta'')}{f(z_i'|\theta')/F(z_i'|\theta')} \geqslant \frac{f(x_i|\theta'')/F(x_i|\theta'')}{f(x_i|\theta')/F(x_i|\theta')},$$

which again follows from log-supermodularity (resp. log-submodularity) of the reverse hazard rate when $m \geqslant k$ (resp. $m \leqslant k$), because $m \geqslant k$ implies $z_i' \geqslant x_i$ (resp. $m \leqslant k$ implies $z_i' \leqslant x_i$). □

Proof of Lemma 1. By independence of the vectors $(\delta_1, \ldots, \delta_k)$ and $(\varepsilon_1, \ldots, \varepsilon_k)$, and by the identical distribution assumption on $\delta_1, \ldots, \delta_k$, it follows that $\delta_{i^*} + \varepsilon_{i^*}$ has the same distribution as $\delta_1 + \varepsilon_{i^*}$. Since δ_1 is independent of $(\varepsilon_1, \ldots, \varepsilon_k)$ and has a logconcave density, the convolution $\eta_{i^*} + \varepsilon_{i^*}$ is less (more) dispersed than $\eta_i + \varepsilon_i$ whenever ε_{i^*} is less (more) dispersed than ε_i, as shown in Theorem 7 of Lewis and Thompson (1981). See also Theorem 3.B.9 in Shaked and Shanthikumar (2007). The conclusion now follows from Theorem 5.2 in Lehmann (1988). □

Proof of Proposition 2. The researcher first corrects for control variables z_{-j} and then estimates θ with OLS. The first step uses (11) to adjust the outcome variable into

$$\hat{y} = y - \sum_{l \neq j} z_l = x\theta + z_j + v.$$

The second step provides the familiar OLS estimate

$$\left(x^T x\right)^{-1} x^T \hat{y} = \left(x^T x\right)^{-1} x^T \left(x\theta + z_j + v\right) = \left(x^T x\right)^{-1} x^T \left(x\theta + x\varepsilon_j + u_j + v\right),$$

where the algebraic manipulations used (11) and (12). This reduces to the claimed expression. □

Proof of Proposition 3. It follows from the analysis above that the evaluator's strategy satisfies the property: for any sample pair $x' \geqslant x$, if a is an optimal choice at x then any optimal choice a' at x' has $a' \geqslant a$. From presample $(x_1, \ldots, x_k)$, selection of the n greatest elements provides a sample vector that dominates any other sample after ranking their elements. The evaluator's strategy is independent of such re-ranking. The posited strategy is thus a best response for the researcher whose utility increases in action. □

Proof of Proposition 4. Since the cost function is convex, it suffices to check that the first two terms in (13) are concave in k. It suffices to take k to be a real number. The first derivative of a^k is $\log(a)\, a^k$ and the second derivative is $(\log(a))^2 a^k$ which is positive when $a \notin \{0, 1\}$. It is easy to see that the first terms are instead constant in k, if the base is zero or one. □

Proof of Proposition 5. (i) The density for k is $\beta^\alpha (k-1)^{\alpha-1} e^{-\beta(k-1)} / \Gamma(\alpha)$ for $k > 1$, where $\alpha \geqslant 1$ is a shape parameter that guarantees logconcavity of this density, and $\beta > 0$ is the rate parameter in the Gamma distribution. Then the distribution of basic noise in the uncertain experiment is

$$\tilde{F}(\varepsilon) = \int_1^\infty e^{-k\exp(-\varepsilon)} \frac{\beta^\alpha}{\Gamma(\alpha)} (k-1)^{\alpha-1} e^{-\beta(k-1)} dk = \frac{e^{-\exp(-\varepsilon)}\beta^\alpha}{(\beta+\exp(-\varepsilon))^\alpha}.$$

Its density is

$$\tilde{f}(\varepsilon) = \frac{e^{-\varepsilon-\exp(-\varepsilon)}\beta^\alpha(\beta+\alpha+\exp(-\varepsilon))}{(\beta+\exp(-\varepsilon))^{\alpha+1}}.$$

The second derivative of $\log\tilde{f}$ is

$$e^{-\varepsilon}\left[\frac{\beta+\alpha}{(\beta+\alpha+e^{-\varepsilon})^2} - \frac{\beta(\alpha+1)}{(\beta+e^{-\varepsilon})^2} - 1\right].$$

This is negative because

$$\begin{aligned} & (\beta+\alpha+e^{-\varepsilon})^2(\beta+e^{-\varepsilon})^2 \\ = \; & (\beta+e^{-\varepsilon})^4 + 2\alpha(\beta+e^{-\varepsilon})^3 + \alpha^2(\beta+e^{-\varepsilon})^2 \\ > \; & \alpha\left[(\beta+e^{-\varepsilon})^2 - \beta(\beta+\alpha+e^{-\varepsilon})^2 - \beta\alpha - \beta 2(\beta+e^{-\varepsilon})\right], \end{aligned}$$

which holds because the only positive term on the right-hand side is exceeded by the third term on the left-hand side when $\alpha > 1$.[42]

(ii) When k is known in the Gumbel case, the noise distribution is shifted up by $\log k$. The unique best response is a cutoff point that also shifts up by $\log k$. The evaluator's welfare is constant in k. Ex ante, before knowing k, then this constant is also the expected welfare. The mixture is worse in the strong sense of Blackwell (1951, 1953). The ex ante welfare cannot be greater than before. It must be lower because the unique best response varied with k. □

References

Allcott, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130(3), 1117–1165.

An, M. Y. (1998): "Logconcavity versus Logconvexity: A Complete Characterization," *Journal of Economic Theory*, 80, 350–369.

[42] When the Gamma distribution is not logconcave, i.e., for $\alpha < 1$, the inequality fails when β and $e^{-\varepsilon}$ are both small. In the limit, as β tends to zero, the inequality implies $(\alpha+e^{-\varepsilon})^2(e^{-\varepsilon})^2 \geqslant \alpha(e^{-\varepsilon})^2$. For any $\alpha < 1$, once ε is sufficiently large, this fails since $\alpha^2 < \alpha$.

Athey, S. (2002): "Monotone Comparative Statics Under Uncertainty," *Quarterly Journal of Economics*, 117(1), 187–223.

Balakrishnan, N., M. Burkschat, E. Cramer, and G. Hofmann (2008): "Fisher Information Based Progressive Censoring Plans," *Computational Statistics and Data Analysis*, 53(2), 366–380.

Banerjee, A. V., S. Chassang, S. Monteiro, and E. Snowberg (2017a): "A Theory of Experimenters," NBER Working Paper No. 23867.

Banerjee, A. V., S. Chassang, and E. Snowberg (2017b): "Decision Theoretic Approaches to Experiment Design and External Validity," in *Handbook of Field Experiments, Volume 1*, ed. by A. V. Banerjee and E. Duflo, North Holland, 141–174.

Berger, V. W. (2005): *Selection Bias and Covariate Imbalances in Randomized Clinical Trials*, Wiley.

Bickel, P. J. and E. L. Lehmann (1979): "Descriptive Statistics for Nonparametric Models. IV," in *Contributions to Statistics, Jaroslav Hájek Memorial Volume*, ed. by J. Jurečková, Academia, Prague.

Blackwell, D. (1951): "Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 93–102.

——— (1953): "Equivalent Comparisons of Experiments," *Annals of Mathematical Statistics*, 24, 265–272.

Blackwell, D. and J. L. Hodges (1957): "Design for the Control of Selection Bias," *Annals of Mathematical Statistics*, 28(2), 449–460.

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012): "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 102(4), 1279–1309.

Dahm, M., P. Gonzàlez, and N. Porteiro (2009): "Trials, Tricks and Transparency: How Disclosure Rules Affect Clinical Knowledge," *Journal of Health Economics*, 28(6), 1141–1153.

Di Tillio, A., M. Ottaviani, and P. N. Sørensen (2017): "Persuasion Bias in Science: Can Economics Help?" *Economic Journal*, 127(605), F266–F304.

Efron, B. (1971): "Forcing a Sequential Experiment to be Balanced," *Biometrika*, 58(3), 403–417.

Felgenhauer, M. and E. Schulte (2014): "Strategic Private Experimentation," *American Economic Journal: Microeconomics*, 6(4), 74–105.

Fishman, M. J. and K. M. Hagerty (1990): "The Optimal Amount of Discretion to Allow in Disclosure," *Quarterly Journal of Economics*, 105(2), 427–444.

Freedman, B. (1987): "Equipoise and the Ethics of Clinical Research," *New England Journal of Medicine*, 317(3), 141–145.

Glaeser, E. L. (2008): "Researcher Incentives and Empirical Methods," in *The Foundations of Positive and Normative Economics: A Hand Book*, ed. by A. Caplin and A. Schotter, New York: Oxford University Press, 300–319.

Goel, P. K. and M. H. DeGroot (1992): "Comparison of Experiments for Selection and Censored Data Models," in *Bayesian Analysis in Statistics and Econometrics. Lecture Notes in Statistics*, ed. by P. K. Goel and N. S. Iyengar, Springer, New York, NY, vol. 75.

Griliches, Z. (1957): "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39, 8–20.

Grossman, S. (1981): "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, 24(3), 461–83.

Hazelton, M. L. (2011): "Assessing Log-concavity of Multivariate Densities," *Statistics and Probability Letters*, 81(1), 121–125.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Henry, E. (2009): "Strategic Disclosure of Research Results: The Cost of Proving Your Honesty," *Economic Journal*, 119(539), 1036–1064.

Henry, E. and M. Ottaviani (2015): "Research and the Approval Process: The Organization of Persuasion," Bocconi mimeo.

Herresthal, C. (2017): "Hidden Testing and Selective Disclosure of Evidence," University of Cambridge mimeo.

Hoffmann, F., R. Inderst, and M. Ottaviani (2014): "Persuasion through Selective Disclosure: Implications for Marketing, Campaigning, and Privacy Regulation," Bocconi mimeo.

Holmström, B. (1999): "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies*, 66(1), 169–182.

Jewitt, I. (2007): "Information Order in Decision and Agency Problems," Oxford mimeo.

Johnson, J. P. and D. P. Myatt (2006): “On the Simple Economics of Advertising, Marketing, and Product Design,” *American Economic Review*, 96(3), 756–784.

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

Karlin, S. and H. Rubin (1956): “The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio,” *Annals of Mathematical Statistics*, 27, 272–299.

Kasy, M. (2016): “Why Experimenters Might not Always Want to Randomize, and What They Could Do Instead,” *Political Analysis*, 24(3), 324–338.

Khaledi, B.-E. and S. Kochar (2000): “On Dispersive Ordering between Order Statistics in One-Sample and Two-Sample Problems,” *Statistics & Probability Letters*, 46(3), 257–261.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983): *Extremes and Related Properties of Random Sequences and Processes*, Springer.

Lehmann, E. L. (1988): “Comparing Location Experiments,” *Annals of Statistics*, 16, 521–533.

Lewis, T. and J. W. Thompson (1981): “Dispersive Distributions, and the Connection between Dispersivity and Strong Unimodality,” *Journal of Applied Probability*, 18, 76–90.

Marshall, A. W. and I. Olkin (2007): *Life Distributions*, Springer.

Milgrom, P. (1981): “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*, 12(2), 380–391.

Müller, S. and K. Rufibach (2008): “On the Max-Domain of Attraction of Distributions with Log-Concave Densities,” *Statistics and Probability Letters*, 78, 1440–1444.

Neyman, J. S. (1923): “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” Translated in *Statistical Science*, 5(4), 465–480.

Persico, N. (2000): “Information Acquisition in Auctions,” *Econometrica*, 68(1), 135–148.

Quah, J. K.-H. and B. Strulovici (2009): “Comparative Statics, Informativeness, and the Interval Dominance Order,” *Econometrica*, 77, 1949–1992.

Rayo, L. and I. Segal (2010): “Optimal Information Disclosure,” *Journal of Political Economy*, 118(5), 949–987.

Rubin, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology*, 66(5), 688–701.

——— (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6(1), 34–58.

Schulz, K. F. (1995): "Subverting Randomization in Controlled Trials," *Journal of the American Medical Association*, 274(18), 1456–1458.

Schulz, K. F., I. Chalmers, R. J. Hayes, and D. G. Altman (1995): "Empirical Evidence of Bias: Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials," *Journal of the American Medical Association*, 273(5), 408–412.

Shaked, M. and J. G. Shanthikumar (2007): *Stochastic Orders*, Springer.

Shaked, M. and Y. L. Tong (1990): "Comparison of Experiments for a Class of Positively Dependent Random Variables," *Canadian Journal of Statistics*, 18, 79–86.

——— (1993): "Comparison of Experiments of Some Multivariate Distributions with a Common Marginal," *Stochastic Inequalities*, 22, 79–86.

Tetenov, A. (2016): "An Economic Theory of Statistical Testing," Bristol mimeo.

Wald, A. (1945): "Sequential Tests of Statistical Hypotheses," *Annals of Mathematical Statistics*, 16(2), 117–186.

B Supplementary Appendix: Additional Extensions

B.1 Other Forms of Selection: Truncated Data

The particular form of selection that is our focus in this paper is but one instance of lack of randomness in empirical or experimental data. Another kind of selection that is often relevant—for instance in the regression contexts discussed in Section 4.2—involves (independent) observations from a *truncated distribution*. Here we review this kind of selection and contrast it with the form of selection analyzed earlier.

Given a random variable with distribution $F(\cdot|\theta)$ and density $f(\cdot|\theta)$ satisfying the MLR property, and given two left-truncation points $-\infty \leqslant a < b < \infty$, define $Y_a := X|X \geqslant a$ and $Y_b := X|X \geqslant b$. Similarly, given two right-truncation points $-\infty < c < d \leqslant \infty$, define $W_c := X|X \leqslant c$ and $W_d := X|X \leqslant d$. Does the evaluator prefer the more left-truncated experiment Y_b or the less left-truncated experiment Y_a? And how do W_c and W_d compare?[43] Using the notion of accuracy, and variants of the arguments used to establish Theorem 2, we obtain a new, simpler proof of the following result, due to Goel and DeGroot (1992).

Theorem 3. *Experiment Y_a is more accurate than Y_b if the hazard rate $f(x|\theta)/[1-F(x|\theta)]$ is log-supermodular. Moreover, W_d is more accurate than W_c if the reverse hazard rate $f(x|\theta)/F(x|\theta)$ is log-supermodular.*

Proof. Consider first the comparison between W_c and W_d. In any state θ, the cumulative distribution function of W_c is $F(w|\theta)/F(c|\theta)$, for $w \leqslant c$. Similarly, the cumulative distribution function of W_d is $F(w|\theta)/F(d|\theta)$, for $w \leqslant d$. Thus, the function ζ_θ such that $\zeta_\theta(W_c)$ has the same distribution as W_d is defined as follows: for every $w \leqslant c$,

$$F(w|\theta)/F(c|\theta) = F(\zeta_\theta(w)|\theta)/F(d|\theta).$$

Now, fixing two states θ' and $\theta'' > \theta'$, we must show that $\zeta_{\theta''}(w) \geqslant \zeta_{\theta'}(w)$ for all $w \leqslant c$. But, by definition of $\zeta_{\theta''}$, we have $F(\zeta_{\theta''}(w)|\theta'')/F(d|\theta'') = F(w|\theta'')/F(c|\theta'')$, so it suffices to show that

$$F(w|\theta'')/F(c|\theta'') \geqslant F(\zeta_{\theta'}(w)|\theta'')/F(d|\theta'').$$

The inequality holds (with equality) in the limit as w increases to the upper bound c, because both sides converge to one. Thus, all we need to prove is that the ratio between the right-hand side and the left-hand side of the inequality increases with w. Taking derivatives, this condition says that

$$\frac{f(w|\theta'')/F(w|\theta'')}{f(w|\theta')/F(w|\theta')} \leqslant \frac{f(\zeta_{\theta'}(w)|\theta'')/F(\zeta_{\theta'}(w)|\theta'')}{f(\zeta_{\theta'}(w)|\theta')/F(\zeta_{\theta'}(w)|\theta')},$$

[43] Here $a = -\infty$ means that $Y_a = X$, so that Y_a is a random draw from $F(\cdot|\theta)$. Similarly for W_d, when $d = \infty$.

which holds precisely when the reverse hazard rate is log-supermodular, given that $\zeta_{\theta'}(w) \geqslant w$ by the fact that W_d first-order stochastically dominates W_c.

Next, consider the comparison between Y_a and Y_b. In any state θ, the cumulative distribution function of Y_a is $[F(y|\theta) - F(a|\theta)]/[1 - F(a|\theta)]$, for $y \geqslant a$. Similarly, the cumulative distribution function of Y_b is $[F(y|\theta) - F(b|\theta)]/[1 - F(b|\theta)]$, for $y \geqslant b$. Thus, the function ζ_θ such that $\zeta_\theta(Y_b)$ has the same distribution as Y_a is defined as follows: for every $y \geqslant b$,

$$[F(y|\theta) - F(b|\theta)]/[1 - F(b|\theta)] = [F(\zeta_\theta(y)|\theta) - F(a|\theta)]/[1 - F(a|\theta)].$$

As before, fixing two states θ' and $\theta'' > \theta'$ we must show that $\zeta_{\theta''}(y) \geqslant \zeta_{\theta'}(y)$ for every $y \geqslant b$, and using the definition of $\zeta_{\theta''}$ it suffices to show that

$$[F(y|\theta'') - F(b|\theta'')]/[1 - F(b|\theta'')] \geqslant [F(\zeta_{\theta'}(y)|\theta'') - F(a|\theta'')]/[1 - F(a|\theta'')].$$

This inequality holds in the limit as y *decreases* to the lower bound b, as both sides converge to one, so we must prove that the ratio between right-hand and left-hand side *decreases* with y, or

$$\frac{f(y|\theta'')/[1 - F(y|\theta'')]}{f(y|\theta')/[1 - F(y|\theta')]} \geqslant \frac{f(\zeta_{\theta'}(y)|\theta'')/[1 - F(\zeta_{\theta'}(y)|\theta'')]}{f(\zeta_{\theta'}(y)|\theta')/[1 - F(\zeta_{\theta'}(y)|\theta')]}.$$

The latter inequality holds when the hazard rate is log-supermodular, given that $\zeta_{\theta'}(y) \leqslant y$ by the fact that Y_b first-order stochastically dominates Y_a. □

The theorem compares one-dimensional experiments. The extension of the result to an arbitrary number of independent observations is immediate. As we have shown earlier, combining conditionally more accurate mutually independent experiments (in this case, more accurate one-dimensional experiments) results in a conditionally more accurate experiment.

Left-truncation type selection bears a resemblance to the kind of selection considered earlier: with both forms of selection, probability mass is moved toward the upper tail of the distribution. In some important cases, however, the welfare consequences of the two types of selection are strikingly different. Consider, for instance, a normal location experiment. In this case the basic noise distribution has both a logconcave hazard rate (see Footnote 28) and a logconcave reverse hazard rate. Moreover, as we have already shown, more selection (of the form analyzed in this paper) benefits the evaluator. However, the more truncated experiment Y_b is worse than the less truncated experiment Y_a. Section 4.2 further discusses this contrast when analyzing (different kinds of) data selection in a regression context.

B.2 Extreme Selection

Adding to our analysis of the impact of selection in one-dimensional location experiments, on the evaluator's welfare, we examine here the effect of extreme selection, with sample size $n = 1$ and presample size $k \to \infty$. We draw on the fundamental result in extreme value theory, which characterizes the limit distribution of the maximum of k i.i.d. random variables, properly normalized for location and scale inflation. Take a basic noise distribution F and suppose that, for some nondegenerate distribution $\bar{F}$ and some sequence of numbers $a_k > 0$ and b_k,

$$F^k\left(b_k + a_k \varepsilon\right) \to \bar{F}\left(\varepsilon\right)$$

for every continuity point ε of $\bar{F}$. The fundamental theorem of extreme value theory says that $\bar{F}$ must belong to one of the following three types: Gumbel, Extreme Weibull or Frechet.[44]

Since the distribution of the noise term is shifted upwards (in the sense of first-order stochastic dominance) as k increases, the location normalization sequence b_k is growing. However, the evaluator can adjust for any translation of the noise distribution without any impact on payoff. The limit impact of selection thus hinges on whether the scale normalization sequence a_k shrinks to zero or not. If $a_k \to 0$, then the noise distribution becomes more and more concentrated around b_k as k grows, providing the evaluator with arbitrarily precise information about the state—the value of the evaluator's problem converges to the full information payoff. If instead we can choose a constant sequence a_k, then we can also choose $a_k = 1$ for all k, and an extremely selected experiment is as accurate as a random experiment based on $\bar{F}$.

It is well known that many familiar distributions are in the domain of attraction of the Gumbel distribution. Specifically, when F is normal—or half-normal, which has the same right tails—then a_k must be decreasing to zero—the scale normalization sequence $a_k = (2\log k)^{-1/2}$ is appropriate in this case—and the limit distribution $\bar{F}$ is the Gumbel distribution. More generally, consider a distribution F in the *exponential power* family, with density

$$f(\varepsilon) = \frac{s}{\Gamma(1/s)} e^{-|\varepsilon|^s},$$

where s is a shape parameter and Γ is the Gamma function. This family includes the Laplace ($s = 1$), normal ($s = 2$), and uniform ($s = \infty$) distributions as special cases. Our next result shows that when the shape parameter s is strictly greater than 1, the scale normalization sequence a_k must be decreasing to zero, and the limiting distribution $\bar{F}$ is the Gumbel distribution.

[44] See e.g. Leadbetter, Lindgren, and Rootzén (1983) for a primer on extreme value theory. As shown in Müller and Rufibach (2008), every logconcave distribution F has a Gumbel or Extreme Weibull limiting distribution $\bar{F}$.

Proposition 6. *Let F be an exponential power distribution with shape parameter $s > 1$. Then $F^k(b_k + a_k\varepsilon) \to e^{-e^{-\varepsilon}}$ for some sequence of constants b_k and $a_k \to 0$. Thus, the value of the evaluator's problem converges to the full information payoff as $k \to \infty$.*

Proof. We show that if $\varepsilon_1, \ldots, \varepsilon_k$ are i.i.d. exponential power with shape $s > 1$, location and scale parameters 0 and 1, then $M_k = \max\langle \varepsilon_1, \ldots, \varepsilon_k \rangle$ satisfies $\Pr(M_k \leqslant a_k\varepsilon + b_k) \to e^{-e^{-\varepsilon}}$ for all ε, where

$$a_k = (s\log k)^{-\frac{s-1}{s}} \qquad \text{and} \qquad b_k = (s\log k)^{1/s} - \frac{\frac{s-1}{s}\log\log k + \log(2\Gamma[1/s])}{(s\log k)^{\frac{s-1}{s}}}.$$

Start by noticing that $f(\varepsilon)/\left(\varepsilon^{b-1}[1-F(\varepsilon)]\right) \to 1$ as $\varepsilon \to \infty$. Fix ε and define y_k for each $k \geqslant 1$ by $1 - F(y_k) = e^{-\varepsilon}/k$, so that

$$\frac{e^{-\varepsilon}}{k}\frac{y_k^{s-1}}{f(y_k)} \to 1 \quad \text{as } k \to \infty. \tag{19}$$

We may assume $y_k > 0$ for all k. Then $f(y_k) = s^{\frac{s-1}{s}} e^{-y_k^s/s}/2\Gamma[1/s]$ and hence, by (19),

$$-\log k - \varepsilon + (s-1)\log y_k - \frac{s-1}{s}\log s + \log(2\Gamma[1/b]) + \frac{y_k^s}{s} \to 0. \tag{20}$$

From (20) we see that $-\log k + (s-1)\log y_k + y_k^s/s = -\log k + o\left(y_k^s/s\right) + y_k^s/s$ converges to a constant. Thus, $-s\log k/u_k^s + o\left(y_k^s/s\right)/(u_k^s/s) + 1$ converges to 0, i.e. $\log y_k = (1/s)(\log s + \log\log k) + o(1)$. Using this fact in (20), we obtain

$$\begin{aligned}\frac{y_k^s}{s} &= \log k + \varepsilon - \frac{s-1}{s}(\log s + \log\log k) + \frac{s-1}{s}\log s - \log(2\Gamma[1/s]) + o(1) \\ &= \log k + \varepsilon - \frac{s-1}{s}\log\log k - \log(2\Gamma[1/s]) + o(1).\end{aligned}$$

Equivalently,

$$\begin{aligned} y_k &= (s\log k)^{1/s}\left[1 + \frac{\varepsilon - \frac{s-1}{s}\log\log k - \log(2\Gamma[1/s])}{\log k} + o\left(\frac{1}{\log k}\right)\right]^{1/s} \\ &= (s\log k)^{1/s}\left[1 + \frac{\varepsilon - \frac{s-1}{s}\log\log k - \log(2\Gamma[1/s])}{s\log k} + o\left(\frac{1}{\log k}\right)\right] \\ &= (s\log k)^{1/s} + \frac{\varepsilon - \frac{s-1}{s}\log\log k - \log(2\Gamma[1/s])}{(s\log k)^{\frac{s-1}{b}}} + o\left(\frac{1}{(\log k)^{\frac{s-1}{s}}}\right) = a_k\varepsilon + b_k + o(a_k).\end{aligned}$$

Thus, $\Pr(M_k \leqslant a_k\varepsilon + b_k + o(a_k)) \to e^{-e^{-\varepsilon}}$, as was to be shown. □

We find the conclusion in Proposition 6 striking, because it is known that when F is the exponential distribution—or the Laplace distribution, since the two distributions have the same right tails—then F^k also converges to the Gumbel distribution, but we can take $a_k = 1$ for each k. (The same normalizing constants work for the generalized exponential distribution defined in (6).) Thus, while extreme selection leads to full information as $k \to \infty$ for any $b > 1$ in the exponential power family, the limit result is very different when $b = 1$. The negative impact of selection in the exponential case discussed earlier is, in this sense, fragile, as any arbitrarily close distribution in the exponential power family reverses the conclusion.[45]

B.3 Uncertain Selection: Non-Monotone Equilibrium

Here we illustrate the construction of a Bayes Nash equilibrium under uncertainty about the presample size k, when the MLR property may fail, and hence the monotone equilibrium from Proposition 3 may not exist. We restrict attention to simple hypothesis testing, and to reduce strategic complexity we assume that $k-1$ follows a Bernoulli distribution: there is probability h that $k = n = 1$ and remaining probability $1-h$ that $k = 2$. For instance, we can think of a heterogeneous population of researchers, some honest (fraction h) and some strategic (fraction $1-h$) as posited for all researchers in the baseline strategic model we presented earlier. At observation x, when the researcher plays maximal selection, the evaluator's likelihood ratio can be written as

$$\frac{f(x-\theta_H)}{f(x-\theta_L)}\left(\hat{h}(x) + \left[1-\hat{h}(x)\right]\frac{F(x-\theta_H)}{F(x-\theta_L)}\right), \tag{21}$$

where $\hat{h}(x) = h/\left[h+(1-h)2F(x-\theta_L)\right]$ may be interpreted as a posterior weight on honesty. This weight is decreasing in x, so weight is gradually shifted towards selection as x grows. When x is very low, there is little correction in the likelihood ratio, as selection is unlikely. When x is very high, there is again little correction because F^2 is very close to F. In the middle, however, the change in weight towards selection can be so powerful that the expression in (21) falls in x.

A (locally) decreasing likelihood ratio has a simple interpretation in the scenario described in this section. Since there is uncertainty not only about the state θ but also about the presample size, the experiment provides information about both, so the net effect on the evaluator's posterior is the result of two forces. On one hand, a higher realization x increases the odds that the state is high, because the MLR property does hold separately in each case, $k = 1$ or $k = 2$. On the other hand, it increases the odds that $k = 2$, and conditionally on $k = 2$ the evaluator may be unwilling to accept even facing the higher realization. In other words, in a region where the likelihood ratio is decreasing, better results indicate greater selection, and are therefore "too good to be true." Figure 7 illustrates this failure of the MLR property where F is the normal distribution.[46]

[45] Of course, as s approaches 1, the convergence to full information becomes slower.

[46] To illustrate a larger effect, here k is either 1 or 5, but the illustration is representative also of the smaller effect

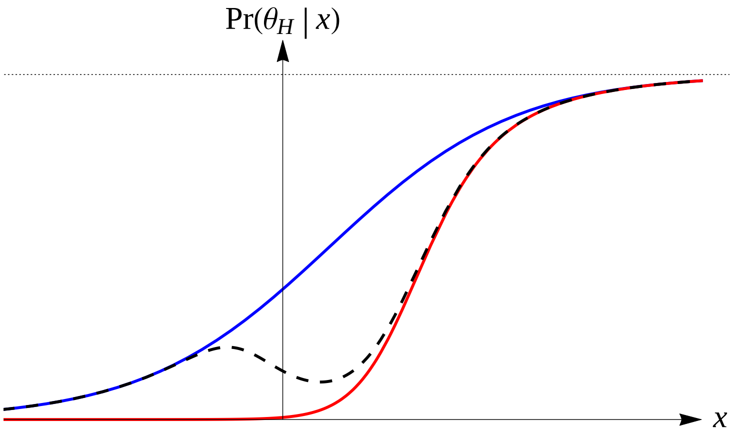

Figure 7: Evaluator's posterior expectation of the state as a function of the experimental observation (dashed curve). With known presample size $k = n = 1$ the expectation would be the blue curve, with known $k = 5$ the red curve.

What if the failure of the MLR property is, in a sense to be made precise, not too large?[47] In this case, we can extend our monotone equilibrium from Proposition 3, as follows. The evaluator continues to accept when the observation exceeds some cutoff point $\bar{x}$. The researcher adopts a new best response that we detail in the proof of the following result. The strategic researcher still obtains approval exactly when the maximal signal exceeds $\bar{x}$. However, when both signals are on the same side of the cutoff point, the researcher with a presample of size $k = 2$ is willing to report the *minimal* signal. Doing this in some instances will modify the evaluator's posterior belief in Bayes Nash equilibrium, justifying acceptance at the cutoff point $\bar{x}$.

Proposition 7. *Consider simple hypothesis testing with $n = 1$ and Bernoulli $k - 1$, so that $k = 1$ with probability h and $k = 2$ with probability $1 - h$. (i) If the failure of the MLR is small, then there exists a continuum of cutoff points $\bar{x}$ consistent with Bayes Nash equilibrium. The evaluator approves when the reported signal exceeds $\bar{x}$. (ii) Comparing the welfare achieved over such equilibria, in the special case of known $k = 2$, the evaluator prefers the equilibrium with cutoff point derived from maximal selection, while the researcher prefers a lower equilibrium cutoff point.*

Proof. To prove part (i), let $g(x|\theta)$ denote the density of the signal realization under maximal selection. For any r in the interior of the support of $g(x|\theta_H)/g(x|\theta_L)$, the evaluator accepts when $g(x|\theta_H)/g(x|\theta_L) \geqslant r$, by (1). Define $\hat{x}(r) = \min\{x|g(x|\theta_H)/g(x|\theta_L) \geqslant r\}$ and $\check{x}(r) =$

obtained when k is 1 or 2.

[47] We define this property in the proof of Proposition 7. It holds, in particular, if there is no failure of the MLR property, e.g., if k is known.

$\max\{x|g(x|\theta_H)/g(x|\theta_L) \leqslant r\}$. By definition, the failure of the MLR is small if, uniformly over all r, the probability $F(\check{x}(r)|\theta_L) - F(\hat{x}(r)|\theta_L)$ is small, and $g(x|\theta_H)/(rg(x|\theta_L))$ is close to 1 on $[\hat{x}(r), \check{x}(r)]$. As illustrated in Figure 7, a small failure typically occurs for a bounded set of r.

Given r, we argue that any cutoff point $\bar{x} \leqslant \hat{x}(r)$ close to $\hat{x}(r)$ is consistent with equilibrium. The construction of the researcher's strategy involves a point $x_a(r) > \check{x}(r)$ in the support of $F(x|\theta_H)$. The researcher who actually has a presample of size $k = 2$ will report the maximal signal except when $x_2 \in [\bar{x}, \check{x}(r)]$ and $x_1 \in [x_a(r), \infty)$. In this special case, the researcher reports the minimal signal, x_2. Intuitively, the researcher seeks to raise the evaluator's posterior on $[\bar{x}, \check{x}(r)]$ by drawing in signal pairs that would otherwise have led to a higher posterior.[48]

Immediately, this is a best response for the $k = 2$ researcher. Where the strategy is modified, both signals exceed $\bar{x}$ and provide acceptance, so there is no loss to reporting the minimum. The researcher with $k = 1$ must report the obtained signal, with no strategic flexibility.

The change in strategy affects the researcher's inference on $[\bar{x}, \check{x}(r)] \cup [x_a(r), \infty)$, and we need to verify that it is optimal to accept when x falls in this set. In equilibrium, the evaluator correctly conjectures the researcher's strategy, updates beliefs with Bayes' rule, and accepts when the likelihood ratio exceeds r.

Consider $x \in [x_a(r), \infty)$. With the proposed strategy, an observation in this interval has density

$$\begin{aligned} \tilde{g}(x|\theta) &= hf(x|\theta) + (1-h)2f(x|\theta)[F(x|\theta) - F(\check{x}(r)|\theta) + F(\bar{x}|\theta)] \\ &= g(x|\theta) - (1-h)2f(x|\theta)[F(\check{x}(r)|\theta) - F(\bar{x}|\theta)]. \end{aligned}$$

The new likelihood ratio is $\tilde{g}(x|\theta_H)/\tilde{g}(x|\theta_L)$. When $\bar{x}$ is near $\hat{x}(r)$ and $F(\check{x}(r)|\theta) - F(\hat{x}(r)|\theta)$ is small, the likelihood ratio is uniformly close to $g(x|\theta_H)/g(x|\theta_L)$. The latter is monotone, and $x_a(r) > \check{x}(r)$ implies that $g(x|\theta_H)/g(x|\theta_L) > r$.

Consider $x \in [\bar{x}, \check{x}(r)]$. With the proposed strategy, an observation in this interval has density

$$\tilde{g}(x|\theta) = hf(x|\theta) + (1-h)2f(x|\theta)[F(x|\theta) + 1 - F(x_a(r)|\theta)].$$

After simple algebra, the inequality $\tilde{g}(x|\theta_H)/\tilde{g}(x|\theta_L) > g(x|\theta_H)/g(x|\theta_L)$ is equivalent to

$$h\left[\frac{1-F(x_a(r)|\theta_H)}{1-F(x_a(r)|\theta_L)} - 1\right] > 2(1-h)F(x|\theta_L)\left[\frac{F(x|\theta_H)}{F(x|\theta_L)} - \frac{1-F(x_a(r)|\theta_H)}{1-F(x_a(r)|\theta_L)}\right].$$

[48] Depending on parameters, there often exists a similar equilibrium with cutoff point $\bar{x} > \check{x}(r)$. Here, the researcher reports the minimum when x_2 lies in a lower interval and $x_1 \in [\hat{x}(r), \bar{x}]$. The lower interval contains signal realizations that on their own have likelihood ratio $f(x|\theta_H)/f(x|\theta_L) > r$, and moving these events may depress the evaluator's posterior belief on $[\hat{x}(r), \bar{x}]$.

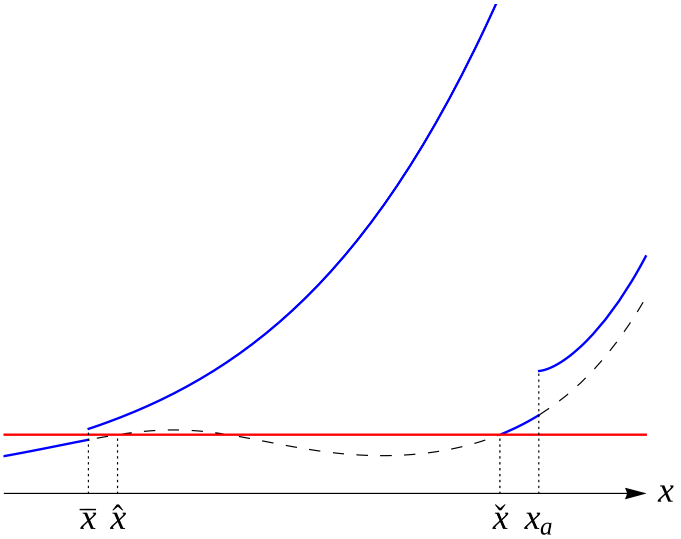

Figure 8: Likelihood ratio in equilibrium of Proposition 7.

Due to the MLR, the left-hand side is positive and the right-hand side is negative, so this is satisfied. The bound is uniform when $\bar{x}$ is near $\hat{x}$ and the failure of MLR is small, so near this limit $\tilde{g}(x|\theta_H)/\tilde{g}(x|\theta_L) > r$.

To prove part (ii) note that, in an equilibrium of the proposed form, there is approval if and only if the researcher's maximal signal exceeds the cutoff point $\bar{x}$. With known $k = 2$, the cutoff point from the equilibrium with maximal selection was optimal for the evaluator. The lower is $\bar{x}$, the more likely is approval, to the benefit of the researcher. □

With the existence of many (a continuum of) equilibria, it may be harder to predict which one is actually played by the two parties. At least, it may seem desirable that the evaluator should keep a monotone acceptance strategy, for normative or positive reasons. Proposition 7 describes such equilibria for small failures of the MLR, for instance due to small h. The construction is illustrated in Figure 8, where the dashed curve represents the likelihood ratio in formula 21, the red line the acceptance threshold r, and the blue curve the equilibrium likelihood ratio. As the researcher with $k = 2$ chooses minimal rather than maximal selection when both presample observations fall in the set $[\bar{x},\check{x}] \cup [x_a,\infty)$, the evaluator's inference becomes more favorable, and monotonicity in the acceptance strategy is restored.

B.4 Local Accuracy

The notion of conditional accuracy provides a sufficient condition that is intuitive and easy to check, and requires no knowledge of the problem parameters: as long as the evaluator's payoff function defines a monotone decision problem, if an experiment Y is conditionally more accurate than experiment X, then the evaluator must prefer Y to X, as shown in Proposition 1. Moreover, as stated in that proposition, in the one-dimensional case considered by Lehmann (1988) (and the following papers based on Lehmann's article) the condition is sharp, for no weaker condition can guarantee that Y is always the preferred experiment—if Y and X are one-dimensional and Y is not more accurate than X, then there exists at least one monotone problem such that the evaluator strictly prefers X to Y.

Despite these convenient features, the accuracy (or conditional accuracy) criterion is often inapplicable—the ordering of experiments in terms of accuracy (or conditional accuracy) is only partial. If the function ζ_θ defined in (7) is not monotone, then the evaluator's preference over X and Y can depend on the specific problem at hand. But suppose now that we do know something about the problem, that is, we are interested in determining the evaluator's preference in some prespecified *subset* of monotone problems. Then, as the class of conceivable problems becomes smaller, more pairs of experiments become comparable.[49] In other words, global monotonicity of the function ζ_θ on the whole set Θ becomes unnecessarily strong a condition, and we may be able to compare experiments by looking at the *local* behavior of the function on a subset of Θ.

To develop the idea a little further, take two location experiments X and Y with respective noise distributions F and G. Since $\zeta_\theta(x) = G^{-1}\big(F(x-\theta)\big)+\theta$, it is easy to see that for every realization x and every $\Delta > 0$ we have

$$\frac{\zeta_\theta(x) - \zeta_{\theta-\Delta}(x)}{\Delta} = 1 - \frac{\zeta_\theta(x+\Delta) - \zeta_\theta(x)}{\Delta}.$$

Besides confirming that Y is more accurate than X if and only if G is less dispersed than F, as proved by Lehmann (1988, Theorem 5.2),[50] the latter equality reveals a correspondence between the shape of the function $\zeta_\theta(x)$ and the shape of the quantile difference $G^{-1}(u) - F^{-1}(u)$. In particular, $\zeta_\theta(x)$ is a bell-shaped (resp. U-shaped) function of θ, in which case we say that Y is *more accurate at the top and less accurate at the bottom* (resp. *more accurate at the bottom and less accurate at the top*) if and only if the quantile difference $G^{-1}(u) - F^{-1}(u)$ is a bell-shaped (resp. U-shaped) function of u.

[49] Lehmann's (1988) concept of accuracy, in turn, shares a similar motivation. By limiting attention to monotone decision problems (and one-dimensional experiments satisfying the MLR property), more pairs of experiments can be ranked than using Blackwell's (1951,1953) notion of sufficiency. Of course, in the extreme case where we focus attention on *one* decision problem, *any* two experiments become comparable, by simply computing which experiment gives the highest expected payoff in that problem. But this computation is, in general, not useful for comparing the same two experiments in a different decision problem. This is precisely why Blackwell's (or Lehmann's) ordering of experiments, despite being partial, is a valuable tool.

[50] See condition (5.6) in Lehmann (1988), an equivalent condition for G to be less dispersed than F.

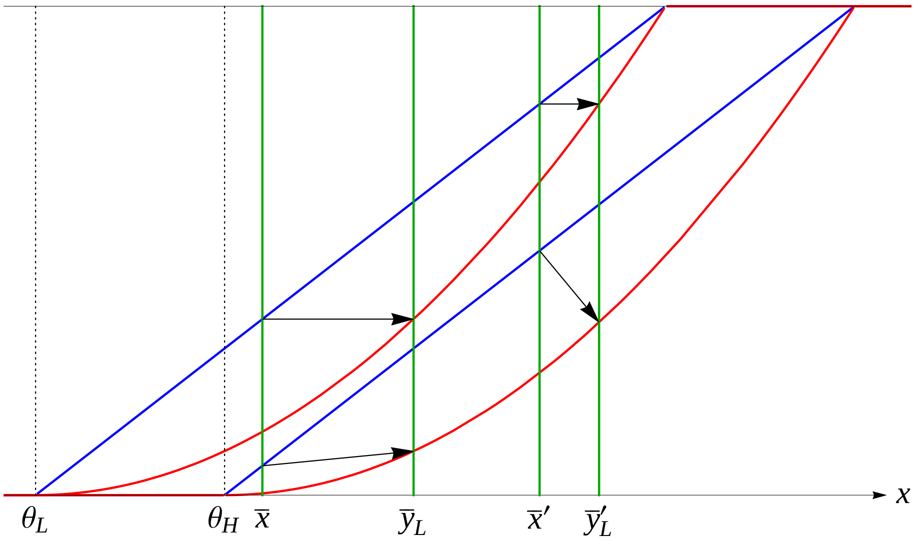

Figure 9: Simple hypothesis testing with uniform basic noise distribution: selection decreases (increases) accuracy for an evaluator strongly favoring acceptance (rejection) a priori.

Now suppose that Y is a selected experiment with presample size k, so that $G = F^k$. Then, by the proof of Theorem 1, the quantile difference is bell-shaped (resp. U-shaped) if and only if the reverse hazard function $-\log F(\varepsilon)$ is logconvex (resp. logconcave) for low values of ε and logconcave (resp. logconvex) for high values of ε. This immediately implies that, in a simple hypothesis testing problem, an evaluator who sets a high cutoff point benefits (resp. loses) from selection, while an evaluator who sets a low cutoff point loses (resp. benefits) from selection. Recall that in simple hypothesis testing with a location experiment with distribution F the evaluator accepts when the likelihood ratio, $f(x-\theta_H)/f(x-\theta_L)$ is at least as large as

$$r = \frac{1-p}{p}\frac{R-\theta_L}{\theta_H - R}.$$

Proposition 8. *Consider an experiment with reverse hazard function $-\log F$ that is first logconvex (logconcave) and then logconcave (logconvex). Then for every $k \geqslant 1$ there exists r_k such that the evaluator prefers F to F^k (resp. F^k to F) for $r \leqslant r_k$ and F^k to F (resp. F to F^k) for $r \geqslant r_k$.*

As illustrated in Figure 9, suppose X corresponds a single random observation drawn from a uniform F (the two blue curves with locations θ_L and θ_H), while Y corresponds to a single selected observation from a presample of size $k = 2$, with distribution F^2 (the two red curves). Suppose that R is sufficiently high, so that the evaluator optimally chooses the high cutoff point $\bar{x}'$ in experiment X. The evaluator stands to gain from switching to Y: choosing cutoff point $\bar{y}'_L$ in Y gives as many false positives, but fewer false negatives. This is because at $\bar{x}'$ the horizontal difference between the selected and the random signal distribution is smaller in the low state than in the high state—the selected experiment is more accurate at the top and less accurate at the bottom. For an evaluator more concerned about false negatives (low R), who was choosing a lower cutoff point like $\bar{x}$ in experiment X, the mechanics of the change to Y work just the opposite way.

Noise drawn from a Laplace distribution, where $F(\varepsilon) = (1/2)e^{\varepsilon}$ for $\varepsilon < 0$ and $F(\varepsilon) = 1 - (1/2)e^{-\varepsilon}$ for $\varepsilon \geqslant 0$, provides our second illustration of Proposition 8. In this case, $-\log F$ is logconcave for $\varepsilon < 0$ and logconvex for $\varepsilon > 0$. Thus, the evaluator's preference for selection is reversed compared to a uniform experiment. Here, the evaluator prefers F^k to F for low values of R, and F to F^k for large values of R. The distribution functions F and F^k are such that the quantile difference $(F^k)^{-1}(u) - F^{-1}(u)$ is U-shaped, and hence so is the function $\zeta_\theta(x)$—the selected experiment is more accurate at the bottom and less accurate at the top.

B.5 Testing for Logconcavity of Reverse Hazard Function

As we have shown in the main text, a basic noise distribution F is such that selection monotonically benefits (or hurts) the evaluator if and only if the basic noise distribution F^k has the same property for every real number $k > 1$. This is because in a one-dimensional location experiment, the reverse hazard functions $-\log F$ $-\log F^k$ only differ by a constant (namely $\log k$). Thus, $-\log F$ is logconcave (logconvex) if and only if $-\log F^k$ is logconcave (logconvex).

This selection-invariance property can be helpful in devising a practical criterion to assess the possible impact of selection in empirical data. Whether selection is known to have occurred or not, empirical outcome distributions with logconvex reverse hazard function should "raise a flag." If selection did occur, then the analyst is bound to having less informative data, even when the analyst is aware of selection and correctly sets the acceptance standard. Instead, a logconcave shape in the data distribution's reverse hazard function indicates that if the analyst does take selection into account, then selection actually results in a more informative experiment.

Suppose an evaluator has obtained data $(x_1, \dots, x_N)$ from N distinct sites, where the observation in each site has been selected, and has computed an estimate $\hat{\theta}$. The noise terms correspond to the residuals $\varepsilon_i = x_i - \hat{\theta}$, which are independent draws from F^k. Then we can use the realized residuals to test $-\log F$ for logconvexity or logconcavity. In the context of field experiments, it is possible to test the same null hypothesis directly on data $(x_1, \dots, x_N)$ under the assumption of homogeneous treatment effect—that is, assuming that θ does not vary across observations, so that distributions coincide up to a constant term.

A first visual assessment of logconvexity or logconcavity can be obtained by plotting the double-log transformation of the empirical cumulative distribution function. Going beyond this suggestive graphical approach, our analysis provides the starting point for the development of empirical tests for logconcavity or logconvexity of the reverse hazard function—for instance, as an extension of Hazelton's (2011) non-parametric test for logconcavity of a density based on a sample of observations.