Persuasion Bias in Science: Can Economics Help?[*]

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

February 28, 2017

Abstract

We investigate the impact of conflicts of interests on randomised controlled trials in a game-theoretic framework. A researcher seeks to persuade an evaluator that the causal effect of a treatment outweighs its cost, to justify acceptance. The researcher can use private information to manipulate the experiment in three alternative ways: (1) sampling subjects based on their treatment effect, (2) assigning subjects to treatment based on their baseline outcome, or (3) selectively reporting experimental outcomes. The resulting biases have different welfare implications: for sufficiently high acceptance cost, in our binary illustration the evaluator loses in cases (1) and (3), but benefits in case (2).

Keywords: Randomised controlled trials; Strategic selection; Welfare

JEL codes: D82, D83, C10, C90

[*] Research funded by the European Research Council through Advanced Grant 29583 (EVALIDEA). We are grateful to Thomas Le Barbanchon, Zach Maniadis and Fabio Tufano for comments, Matteo Camboni and Nicol Rainy Brown for research assistance and participants at seminars in Milan, at the 2015 Conference on Information Aggregation in Bonn and at the 2015 Southampton Conference on the Credibility of Empirical Research for 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

Modern science has made great advances thanks to careful experimentation and statistical analysis of data. But, as an increasing amount of data becomes available and more sophisticated statistical methods are developed, there is growing concern that researchers may be in a better position to select data and use techniques that give the results they desire.[1] Such manipulation endangers the credibility of research.[2] This paper investigates the impact of conflicts of interests when the collection and analysis of empirical evidence is modelled in a game-theoretic framework.

The point of departure of our analysis is the randomised controlled trial (RCT), the gold standard procedure to estimate the causal effect of an intervention.[3] According to an ideal RCT experiment, a researcher should proceed as follows:

1. The researcher takes a representative (e.g. random) sample of the population of interest.
2. Out of the sample, the researcher then forms two groups through random allocation:
   (a) One group is exposed to the intervention (treatment group).
   (b) The other group receives no treatment (control group).
3. The two groups are followed up for a specified period of time, and the effects of the intervention are found by comparing the two groups in terms of a set of outcomes defined at the outset.

The procedure aims at minimising selection bias and maximising applicability of the experimental results. Given that the two groups are handled identically apart from the intervention received, differences in outcomes are correctly attributed to the intervention (internal validity). The representativeness of the sample then makes the estimate valid outside the sample (external validity). To be sure, a researcher in pursuit of the truth has no incentive to deviate from the ideal RCT. In reality, however, researchers are specialised agents who are subject to their own set of incentives, and their goals often conflict with the interests of final research users. For example, clinical trials to test the safety and efficacy of a new drug are typically sponsored by pharmaceutical companies that have vested interests in drug approval. Given the researcher's direct role in the study, strict adherence to the standard is at best an idealisation.

Historically, experimental methodology developed to curtail the researcher's own ability to bias research results—we discuss this in detail in the next section. The standard decision-theoretic approach to statistical inference, however, does not model the researcher's motives, and thus the welfare impact of selection bias remains unqualified. To account for conflicts of interest, Di Tillio,

[1] See for example the leader on the July 25, 2015 issue of *The Economist* titled 'Drug testing: Trials and errors'.

[2] See, for example, Simonsohn, Nelson and Simmons (2014), and Head et al. (2015) for a recent account of p-hacking across sciences.

[3] For accessible introductions to experimental methodology and biases see Jadad and Enkin (2007), Torgerson and Torgerson (2008) and Weisberg (2010).

Ottaviani and Sørensen (2017) cast statistical inference as a strategic problem, by superimposing a game-theoretic structure over Rubin's (1978) Bayesian rendition of Neyman's (1923) potential outcomes framework. In our model, a researcher provides experimental evidence to an evaluator who must then decide whether or not to accept a hypothesis—in the drug industry, whether to approve a new drug.[4] The two players have different objectives: the researcher is biased towards acceptance while the evaluator seeks to avoid both type I errors (false positives) and type II errors (false negatives). The model thus provides a tool to assess the welfare implications of strategic selection as occurring in equilibrium.

In this paper we study variations of a simplified version of Di Tillio, Ottaviani and Sørensen's (2017) model to discuss the welfare impact of strategic deviations at the three critical junctures defining an ideal RCT:

Selective Sampling. First, the researcher may non-randomly select the sample, which can undermine the external validity of the experiment. An experiment is not externally valid if it is based on a sample that is not representative of the general population to which the treatment is intended to be rolled out. External validity has long been recognised as critical in medicine, see for example Rothwell (2005) for discussion and references. Allcott (2015) provides extensive evidence of how the estimated average treatment effects vary across a sequence of 111 RCTs on energy savings carried out by a sponsoring company in partnership with local energy utilities across the US. Initial RCTs tend to be selectively implemented in high impact sites, with impact declining at later sites. As Alcott shows, in spite of the large total sample of the trials covering over 8 million Americans, it is difficult to make a reliable inference of the average treatment effect for the general population of interest.

Selective Assignment. Second, the researcher may use private information on subjects' characteristics, e.g. their outcome when not receiving the treatment, to non-randomly select subjects into treatment, leading to a violation of internal validity. If the treatment group systematically differs from the control group, the evaluator cannot interpret presented evidence as the outcome of a properly conducted RCT on the sample population. To avoid this problem, carefully designed experiments set out strict procedures for obtaining a random assignment and concealing the allocation from researchers. In a medical context, for example, clinicians administering the drug should not know whether a patient receives the test drug or the placebo. However, in practice it is difficult to achieve careful randomisation and to conceal the allocation from the researcher. As documented by Schulz (1995) in a classic piece, clinical investigators often find it difficult to maintain impartiality and find many ways to subvert randomisation, for example by holding up sealed envelopes to lights in order to dictate the assignment of the next patient. According to Schulz at al. (1995), reported treatment effects

[4] By focusing on the pattern that determines which observations are missing, Neyman-Rubin's potential outcomes framework has proved useful for the analysis not only of experimental data but also of observational data; see Angrist and Pischke (2009) and Imbens and Rubin (2015). As a first stab at the problem, our analysis considers experimental inference, but we expect our methodology and insights to extend to observational data.

in studies with inadequate allocation concealment are much (40%) larger than those from properly randomised trials.

Selective Reporting. Third, both external and internal validity are challenged when the researcher decides post hoc which results to report after conducting multiple experiments. Essentially, in this case the researcher selectively discloses evidence. In this paper we do not explicitly consider the related case in which the researcher reports a subset of the outcome variables or does not properly adjust for multiple hypothesis testing—a problem that has attracted large attention in the statistics literature.[5]

In all three cases, the researcher's ability to manipulate crucially depends on private information held prior to the experiment—in which sites the treatment is larger, what baseline characteristics lead to more favourable results, or which result is more favourable. Accordingly, all three model variants feature private information on the side of the researcher.

The different manipulative actions that the researcher takes have a direct impact on the production of evidence. In turn, the evaluator attempts to correct for any biases introduced, and thus adjusts the inference. The reaction of the evaluator constitutes an additional, indirect channel through which manipulation impacts the outcome of research. In our game-theoretic approach, we characterise the equilibrium that results when the researcher manipulates, taking into account the evaluator's inference, and the evaluator's inference correctly anticipates the manipulation.[6]

How is the welfare of researcher and evaluator affected by manipulation? Overall, we find that informed manipulation undertaken by the researcher can generally harm or benefit the two parties, depending on parameters. Holding fixed the evaluator's reaction, clearly the researcher benefits from manipulation. We characterise when the impact of manipulation in equilibrium—once the evaluator adjusts the inference—ends up being negative for the researcher. Turning to the evaluator, the possibility arises that the manipulated experiment contains more information than in the benchmark without manipulation, given that manipulation is based on private information held by the researcher. On the other hand, the researcher uses this information to increase the probability of acceptance, and thus not directly in the interest of the evaluator.

Our results are based on an all-binary specification of Di Tillio, Ottaviani and Sørensen's (2017) model.[7] In the realistic case where the evaluator's opportunity cost of acceptance is high, our model predicts that a rational evaluator stands to gain from manipulation that challenges internal validity, but not from manipulation that challenges external validity. Intuitively, information used to differentiate the treatment and control groups can reduce the noise of confounding variables when comparing the outcomes of the two groups. Instead, a less demanding evaluator, who only rejects when the evidence is strongly negative, stands to gain from selective sampling—as such

[5] See, for example, Rosenthal (1979), Denton (1985), Benjamini and Hochberg (1995) and Ioannidis (2005). On meta-analyses and replication see Ioannidis, Stanley and Doucouliagos (2015) and Maniadis, Tufano and List (2014) and (2015).

[6] We also consider the effect of manipulation on the unsuspecting evaluator.

[7] See also footnote 39 in Section 7.

manipulation renders bad news particularly convincing—but not from selective assignment. In some situations, both the researcher and the evaluator suffer from the researcher's ability to manipulate. When such a credibility crisis arises, both parties would gain from strict procedures which credibly prevent the researcher from manipulating evidence.

The paper proceeds by first reviewing how the RCT methodology historically developed in response to conflicts of interest in scientific disputes. Section 3 introduces our game-theoretic framework, initially focusing on the classical benchmark without conflicts of interest. Section 4 analyses external validity problems due to selective sampling. Section 5 studies challenges to internal validity due to selective assignment. Given that internal validity problems rest on differences among the treatment and control group, for the case of selective assignment Section 5.2 discusses the informative role played by the control group in an RCT. Section 6 turns to selective reporting of outcomes. Section 7 concludes by framing our contribution to the broader literature on persuasion and selective disclosure in economics.

2 Incentives in Experimentation: A Historical Perspective

Considering historical sources and developing a thesis put forward also by Chalmers (2001), this section traces the evolution of experimentation techniques to efforts to control biases strategically introduced by researchers.[8] Early accounts of controlled statistical experiments were couched in adversarial settings where contenders wanted their point of view to prevail. To prove one's point convincingly against the contrasting view of skeptical opponents, it became necessary to find a convincing and fair method to establish causal inference. Randomised assignment of experimental subjects then naturally emerged as a fair way to allocate possibly heterogeneous subjects to different treatments.

The advent of the pharmaceutical industry created a strong advocate for demonstrating favourable results of drugs, potentially leading to selective assignment of experimental subjects that were more likely to show positive outcomes. As health care provision became more centralised, the drug approval process and the modern clinical trial were perfected. Randomisation in the assignment to treatment and control was adopted as a commitment device to avoid selection bias.

In practice, however, adoption of RCT does not completely eliminate the scope of manipulation. Trial patients are often drawn from a selected subpopulation, questioning external validity of the results when the treatment is applied to the target population. Many clinical trials are not properly controlled and randomisation is often subverted, challenging internal validity. Researchers often have some freedom in choice in selecting the outcomes to report.

[8] The James Lind Library (http://www.jameslindlibrary.org) is an excellent source of many original records on the history of experimental methodology. Our discussion of the genesis of the potential outcomes framework also draws on Rubin (2005) and Imbens and Rubin (2015).

Genesis of Controlled Experiments. The use of experiments in decision making is at least as old as the Bible. To prove his diet superior to the diet provided by King Nebuchadnezzar, Daniel proposed the following scheme:

> 11 Then said Daniel to Melzar, whom the prince of the eunuchs had set over Daniel, Hananiah, Mishael, and Azariah, 12 Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink. 13 Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants. 14 So he consented to them in this matter, and proved them ten days. 15 And at the end of ten days their countenances appeared fairer and fatter in flesh than all the children which did eat the portion of the king's meat. 16 Thus Melzar took away the portion of their meat, and the wine that they should drink; and gave them pulse.
> — Daniel 1:11-16

Arguably, Daniel's test could be considered the first controlled experiment. The design captures the gist of the idea of estimating a causal effect by comparing the outcome of treated and untreated units after a pre-specified period. However, there was wide scope for manipulation in Daniel's experiment, and Daniel's objective when proposing the experiment was to prove that his favourite diet (pulse to eat and water to drink) was healthier than the royal diet (meat and sophisticated wine).[9] Daniel selected the experimental units and included himself in the subject pool. In addition, the outcomes were only loosely defined in advance in terms of countenance, so the final evaluation could only be highly subjective. There is magical aura in the results, with *all* subjects displaying a positive treatment effect.

The outcomes claimed in this and other early experiments are rather striking and display "incredible certitude" in the results, borrowing a term from Manski (2013). According to these and other early accounts,[10] experimental techniques seem to originate more from the desire to show the superiority of one's favourite treatment than from the human innate need to understand the world.

Randomisation as Dispute Resolution. We have to wait until the 17th century for consideration to be given to the method that should be used for assigning subjects to different treatments. Jan Baptiste van Helmont suggested randomising the allocation of the batch of patients in the following passage:

[9] We presume that Daniel's experiment was not crucial for the United Nations' decision to declare 2016 the UN International Year of Pulses.

[10] In particular, see the record from the James Lind Library on Leonardo Fioravanti's defence in 1573. To defend himself from the charge of having prescribed treatment resulting in the death of some of his patients, Fioravanti asked to be judged by the following clinical trial:

> . . . that there be consigned to me alone twenty or twenty-five sick people with diverse ailments, and an equal number with the same infirmities to all the physicians of Milan, and if I don't cure mine faster and better than they do theirs, I am willing to be banished forever from this city. . .

> Let us take from the itinerants' hospitals, from the camps or from elsewhere 200 or 500 poor people with fevers, pleurisy etc. and divide them in two: let us cast lots so that one half of them fall to me and the other half to you. I shall cure them without blood-letting or perceptible purging, you will do so according to your knowledge (nor do I even hold you to your boast of abstaining from phlebotomy or purging) and we shall see how many funerals each of us will have: the outcome of the contest shall be the reward of 300 florins deposited by each of us. Thus shall your business be concluded.
> — Jan Baptiste van Helmont (1662)

Note that van Helmont's experiment is organised as a contest pitting two competing therapeutic methods. Again, the proponent of the experiment claims superiority of his treatment and challenges the supporters of the established treatment. The question arises of how to allocate the patients to the two groups. Given that the contenders have strict economic incentives to show that their treatment is superior, both contenders want to enrol the healthier patients in their camp. Randomisation by casting lots is the natural way to obtain a fair allocation, with a lacklustre tradition in all matters of dispute resolution. According to this interpretation, the rationale for randomisation in clinical trials is fairness.[11] While modern randomisation is done subject by subject, here randomisation is for the entire batch of half of the subject pool. A fortunate by-product of randomisation is that we can attribute the differences in outcomes to differences in the effectiveness of the treatments rather than to confounding factors—whether and how well this point was understood is unclear. Also, we have no indication whether this experiment, like the others reported above, really took place.

The idea of comparing similar groups gained momentum and started being practised in the 18th and 19th centuries, leading to important scientific discoveries. This era is epitomised by James Lind's test to find the best treatment for sailors affected by scurvy. In his pioneering 1747 experiment, James Lind compared the effectiveness of six treatments by administering each treatment to two scorbutic sailors. Lind emphasised that the patients were in a similar initial state and so this was a like-with-like comparison. The patients who were given limes and oranges recovered well and quickly. However, it took some time for these results to become accepted and the treatment to be more widely adopted. It is natural to wonder whether a sample larger than two per treatment would have been more convincing.

Alternation. As controlled testing gained increased acceptance, reporting of results became more detailed. More details regarding the construction of the sample were reported by the original investigators. In his 1816 doctoral thesis at the University of Edinburgh, Alexander Hamilton claimed to have assessed the effects of bloodletting in a large sample of 366 sick soldiers as follows:

> It had been so arranged, that this number was admitted, alternately, in such a manner that each of us had one third of the whole. The sick were indiscriminately received,

[11] In addition to carrying scales and a sword, Iustitia, the female goddess of justice of ancient Rome, was typically depicted wearing a blindfold. Fortuna, the goddess of chance and personification of luck, was also depicted veiled or blind.

> and were attended as nearly as possible with the same care and accommodated with the same comforts. One third of the whole were soldiers of the 61st Regiment, the remainder of my own (the 42nd Regiment). Neither Mr. Anderson nor I ever once employed the lancet. He lost two, I four cases; whilst out of the other third (treated with bloodletting by the third surgeon) thirty five patients died.
> — Alexander Hamilton (1816), as quoted by Chalmers (2001)

Like randomisation, alternation can be seen as a fair method of sharing patients among the three experimenting surgeons. Naturally, selection to treatment through alternation guards against selection bias—and it is now done patient by patient. It also ensures some minimal degree of external validity, in so far as the outcomes obtained in one of the groups are representative of the outcomes that would have obtained in the other groups. Hamilton is also well aware that all patients in the same treatment groups should be treated in the same way. However, the treatment is rather loosely defined and the outcome variable also is not clearly set in advance. Be it as it may, Chalmers (2001) reports that Hamilton's biographer judges this entire account to be 'a fabrication, made up for the purpose of obtaining a degree and impressing his readers'.

Randomisation for Valid Testing of Significance. Zooming forward to the development of the theory and the modern practice of experimentation, we must turn to Ronald A. Fisher and the English School in the 1920s. In Fisher's construction, anticipated in Fisher (1925, 1926) and more fully developed in (1936), randomisation is a key ingredient for valid inference. The errors are correctly estimated and the significance test is validated only with proper randomisation:[12]

> One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random, so that no distinction can creep in between pairs of plots treated alike and pairs treated differently; in such a case an estimate of error, derived in the usual way from the variation of sets of plots treated alike, may be applied to test the significance of the observed difference between the averages of plots treated differently. The estimate of error is valid, because, if we imagine a large number of different results obtained by different random arrangements, the ratio of the real to the estimated error, calculated afresh for each of these arrangements, will be actually distributed in the theoretical distribution by which the significance of the result is tested.
> — Fisher (1926), p. 506–507.

Working in parallel to the English School, Neyman (1923) introduced the notation for the potential outcomes framework and analysed inference in a completely randomised experiment. See Imbens and Rubin (2015) for a discussion of the connection between the ideas of Fisher and those of Neyman as well as for an overview of applications of the potential outcomes framework to the analysis of observational data—the Rubin causal model.

[12] See also Fisher (1936) for a critical appraisal of Mendel's statistical evidence.

Allocation Concealment. While the statistical theory of experiments was developed in the agricultural arena, the practice of medical trials was also making great strides. As documented in Kaptchuk's (1989) account, the first instances of treatment randomisation combined with blind assessments date back to the late eighteenth century and were used as part of anti-fraud campaigns to debunk dubious and unconventional medical practices such as animal magnetism. The first record of randomisation seems to be based on a report published in 1785 by a commission of leading scientists and physicians appointed by the King of France and headed by Benjamin Franklin, the American inventor then ambassador to France.

The modern clinical trial evolved in the first half of the twentieth century with the advent of potentially powerful but also dangerous new therapeutic drugs. Particularly momentous were two landmark trials conducted in the UK in the 1940s: the Medical Research Council (1944) clinical trial of the effect of patulin on the common cold and the Medical Research Council (1948) trial of streptomycin for the treatment of pulmonary tuberculosis. These two controlled trials pioneered the practice of concealing the allocation to treatment and control from the patients and from the clinicians at the moment of assignment; see the account by D'Arcy Hart (1999), the secretary of the two committees running these trials.[13]

The patulin trial was placebo controlled, with the assignment to control and treatment done by alternation. The designers recognised the importance of preventing foreknowledge of allocations from the clinicians in charge of admitting patients to the study.[14] To "muddle people up" two treatment (patulin) groups and two control (placebo) groups were used with allocation of patients to these four groups dictated by strict rotation.

The time was ripe for the next step—randomisation with concealed assignment. In the statistical design of the 1948 streptomycin trial masterminded by Austin Bradford Hill, the key rationale for randomising the assignment was to facilitate concealment. As reported in the anonymous editorial of the British Medical Journal:

> Then to ensure that the patient did conform to the features laid down the Committee set up a selection panel. This panel conceivably might have been influenced in selecting or rejecting a patient if it had known beforehand whether the patient was to be allocated to the streptomycin or to the controlled group—e.g., if alternate patients had been taken. It was relieved of any such worries by an ingenious system of sealed envelopes. Once a patient had been accepted an appropriate numbered envelope was opened, and not till then was the patient's group revealed. The allocation to "S" or "C" in this form had been

[13] See also Chalmers (1999). These trials were also important for establishing a set of good practices in conducting the clinical trials and reporting the results. The aim of the studies were clearly set out in advance. For example, in the 1948 Medical Research Council streptomycin trial the aim was to measure the effect of streptomycin upon respiratory tuberculosis. The initial condition was more precisely defined by restricting the trial to "acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proved, unsuitable for collapse therapy, age group 15 to 25 (later extended to 30)" (p. 770). Thus, the questions asked of the trial were deliberately limited, and these "closely defined features were considered indispensable, for it was realised that no two patients have an identical form of the disease, and it was desired to eliminate as many of the obvious variations as possible" (p. 770).

[14] See also Chalmers and Clarke (2004).

> made at random by the statistician. It is instructive also to see how close an equality of group characteristics this statistical method produced.
> — British Medical Journal (1948), p. 792.

In the 1955 edition of his leading textbook on medical statistics, Bradford Hill reiterates that the reason for randomising the allocation and concealing the assignment from clinicians was to reduce the risk of manipulation.

> In many trials this allocation has been successfully made by putting patients, as they present themselves, alternately into the treatment and control groups. Such a method may, however, be insufficiently random if the admission or non-admission of a case to the trial turns upon a difficult assessment of the patient and if the clinician involved knows whether the patient, if accepted, will pass to the treatment or control group. By such knowledge he may be biased, consciously or unconsciously, in his acceptance or rejection; or through fear of being biased, his judgment may be influenced. The latter can be just as important a source of error as the former but is more often overlooked. For this reason, it is better to avoid the alternating method and to adopt the use of random sampling numbers; in addition, the allocation of the patient to treatment or control should be unknown to the clinician until after he has made his decision upon the patient's admission. Thus he can proceed to that decision—admission or rejection—without any fear of bias.
> — Bradford Hill (1955), p. 239–240.

Subversion of Randomisation. Note that the assignment can also be concealed from the clinicians in charge of selecting patients for the trials also with alternation. However, the risk that allocation concealment falls through is high with alternation because then it is enough to learn the assignment of one patient to discover the entire allocation. Randomisation is generally less susceptible to unravelling than alternation. However, common implementations of randomisation, even if combined with allocation concealment before assignment, are not immune from the risk that the assignment schedule becomes partially predictable:

1. The risk of selection bias can be minimised by a simple random allocation of patients to treatment and control. However, when each patient has equal chance of being assigned to treatment versus control, it is likely that the final allocation results in an unequal number of individuals in the two arms of the study. This treatment imbalance decreases the power to detect statistically significant differences between groups. To eliminate or reduce the probability of treatment imbalances, restricted randomisation procedures are commonly adopted. For example, consider the common practice of block randomisation according to which a fixed number of patients are assigned to treatment and control, with a random assignment within the block. Suppose that the random assignment has been properly concealed before assignment, but that clinicians admitting patients to the trial learn the treatment to which the

previous patients have been assigned. If the treatment assignments become known after allocation, some future assignments become partially predictable, regardless of the effectiveness of allocation concealment. For example, the assignment of the last patient in the block is always predictable. More generally, the assignment of a patient late in the block can be partly predicted whenever there is some assignment imbalance in the previous patients.[15]

2. A second practical instance in which clinicians may be able to predict the assignment involves direct deciphering of the concealed allocation schedule by the clinicians in charge of screening patients. Schulz (1995) and Schulz et al. (1995) present evidence that researchers often take advantage of inadequate allocation concealment schemes to subvert randomisation. For example, in schemes involving envelopes, researchers may open envelopes in advance of the allocation or hold envelopes up to light sources to uncover the allocation codes. Researchers then assign to treatment patients that are more likely to display favourable outcomes.

Knowledge of the next assignment can lead clinicians to exclude certain patients based on their prognosis, or to direct a participant's entry into the trial until the next appropriate assignment occurs. Avoidance of such bias depends on the prevention of foreknowledge of treatment assignment. Comparing blind randomised trials, unblinded randomised trials and non-randomised trials, Chalmers et al. (1983) found evidence that the treatment allocation was biased unless the trials were randomised and blind. Also, on average non-random trials had a general imbalance of prognostic factors in favour of the experimental group. They concluded that knowledge of the treatment to be allocated had interfered with the randomisation process.

3 Statistical Environment

Setup. We introduce a stylised model to illustrate our key insights on persuasion bias. In our model, a game is played by a researcher and an evaluator. The researcher sets up an experiment. The evaluator observes the experimental outcome and decides whether to grant the researcher a desired acceptance.[16] The aim of the experiment is to estimate the effect of a treatment (e.g., a new drug or proposed policy) on an outcome of interest (e.g., health or income). The evaluator accepts if the evidence is sufficiently favourable for the treatment. The researcher, instead, always benefits from acceptance.

We model the experiment as a randomised controlled trial in the potential outcomes framework pioneered by Neyman (1923) and Rubin (1974). In this framework, each individual i is characterised by two potential outcomes, at most one of which can be observed. The first potential outcome is the *baseline outcome* (e.g., the individual's health when not taking the new drug), denoted by $Y_i(0)$,

[15] See Blackwell and Hodges (1957). Efron (1971) had developed an alternative "bias coined design" which does not achieve perfect balance between the number of patients in treatment and control. See Rosenberger and Lachin (2002) and Berger (2005) for extensive treatments.

[16] The evaluator's decision captures the essentials of hypothesis testing.

which can only be observed if i is not treated. If instead individual i is treated, then only the other potential outcome, the *treated outcome*, denoted by $Y_i(1)$, can be observed. The treatment effect on individual i is the difference between treated and baseline outcome, $\beta_i = Y_i(1) - Y_i(0)$.

To conduct the experiment, the researcher selects a sample of individuals from the population and divides them into a treatment group and a control group. The experiment results in a list of baseline outcomes for individuals in the control group, and a list of treated outcomes for individuals in the treatment group. A statistical model governs the inference from the observed outcomes in the sample to the potential outcomes in the overall population.

The opportunity to bias the trial is due to private information of the researcher. An external validity problem arises if the researcher can select individuals to enter the trial, or be allocated to the treatment group, based on information about their treatment effect. The population in the experiment is then not representative of the population at large. An internal validity problem arises when the researcher can assign individuals to treatment or control based on information about their baseline outcome. Finally, the researcher may selectively report the experimental outcomes. As we explain later, selective reporting of outcomes poses a challenge to both internal and external validity.

A key simplifying assumption is that we reduce the experiment to be conducted on only two subjects, one treated and one untreated. The essence remains that the experimental outcome carries information about the treatment effect, but this information is imperfect because the baseline outcome of the treated individual may be different from the baseline outcome of the untreated individual. Although simplified, the model allows us to illustrate the key mechanisms in research bias, and to evaluate how this bias affects researcher and evaluator.

Distributions. Let us now turn to describe the details of the statistical model. The individuals in a population N are located in two sites, say Milan and Copenhagen.[17] For simplicity, we assume homogeneous treatment effects within sites: $\beta_i = \beta_M$ if individual i resides in Milan, while $\beta_i = \beta_C$ if i resides in Copenhagen.[18] To simplify some calculations, we conveniently assume that the populations in the two sites have the same size, so that the *average treatment effect* is given by

$$\beta_{ATE} = \frac{\beta_M + \beta_C}{2}.$$

Neither the researcher nor the evaluator initially knows the baseline outcome or treatment effect of any individual, which are therefore random quantities at the outset. Specifically, we assume that baseline outcomes are Bernoulli random variables that are i.i.d. across individuals, with

$$\Pr(Y_i(0) = 1) = p$$

[17] Even though we do not need to make explicit assumptions on the overall number of individuals in the population, it is convenient to think of N as a large set of individuals.

[18] A site can be interpreted more generally as a characteristic of the population that is known to have an impact on the treatment effect.

for each individual i, while treatment effects are Bernoulli random variables that are i.i.d. across sites, with

$$\Pr(\beta_M = 1) = \Pr(\beta_C = 1) = q.^{19}$$

Finally, we assume independence among treatment effects and baseline outcomes: the random variables $\beta_M, \beta_C, (Y_i(0))_{i\in N}$ are mutually independent.

Table 1 illustrates the distribution over the 16 possible combinations of baseline outcomes and treatment effects of two random individuals i and j. The sixth and seventh columns in the table represent the treated outcomes of i and j under the assumption that i lives in Milan and j lives in Copenhagen.

The payoff to the researcher is 1 if the proposal is accepted and 0 if it is rejected. In the latter case, the payoff to the evaluator is also 0, while in case of acceptance, it is $\beta_{ATE} - k$, where $k \in [0,1]$ is the cost of acceptance. Higher values of the parameter k represent a more demanding evaluator. If k is high, acceptance requires very convincing evidence in favour of the treatment. If low, acceptance is granted except when very unfavourable evidence is presented.

Both players maximise expected utility, and we study non-cooperative behaviour.

To analyse the evaluator's and researcher's payoff, it is useful to refer to the cumulative distribution function (henceforth CDF) of the conditional expectation of the average treatment effect. Based on some observed evidence υ, the rational evaluator forms a posterior distribution over β_{ATE}. Let us denote by π the random variable which is this posterior expectation, i.e., $\pi = \mathrm{E}(\beta_{ATE}|\upsilon)$. Let F denote its CDF. The evaluator's payoff from acceptance is $\pi - k$. The rational evaluator accepts when $\pi \geq k$. Hence the researcher's expected payoff is $1 - F(k)$. Using integration by parts, the expected payoff of the evaluator is

$$\int_k^1 (\pi - k)\, dF(\pi) = \int_k^1 (1 - F(\pi))\, d\pi.$$

This means that we can read the evaluator's payoff from the plot of the CDF by considering the area above the curve.

Non-Manipulation Benchmark. We begin our analysis by considering a non-manipulated experiment. The researcher selects a random site, say, site M. A random individual i from the site is assigned to treatment, $t = i$. If the experiment so requires, the researcher also selects from site M another random individual, $c = j$, to act as control. Write $Y_t(1) = Y_i(0) + \beta_M$ for the outcome of the treated individual and $Y_c(0) = Y_j(0)$ for the outcome of the untreated individual.

Actually, the control is uninformative in this experiment, and we can safely ignore it. Here is why. The evaluator bases the acceptance decision on inference about β_{ATE}. But $Y_c(0) = Y_j(0)$ is independent of β_{ATE} and independent of $Y_i(0)$.[20] The observation of $Y_t(1)$ is therefore a sufficient

[19] The assumption that all our variables are Bernoulli distributed simplifies the exposition.

[20] A control observation would be useful if instead the distribution of the baseline outcome were unknown, in which case $Y_j(0)$ would be informative about $Y_i(0)$.

β_M	$Y_i(0)$	β_C	$Y_j(0)$	β_{ATE}	$Y_i(1)$	$Y_j(1)$	Probability
0	0	0	0	0	0	0	$(1-q)(1-p)(1-q)(1-p)$
0	0	0	1	0	0	1	$(1-q)(1-p)(1-q)p$
0	0	1	0	$\frac{1}{2}$	0	1	$(1-q)(1-p)q(1-p)$
0	0	1	1	$\frac{1}{2}$	0	2	$(1-q)(1-p)qp$
0	1	0	0	0	1	0	$(1-q)p(1-q)(1-p)$
0	1	0	1	0	1	1	$(1-q)p(1-q)p$
0	1	1	0	$\frac{1}{2}$	1	1	$(1-q)pq(1-p)$
0	1	1	1	$\frac{1}{2}$	1	2	$(1-q)pqp$
1	0	0	0	$\frac{1}{2}$	1	0	$q(1-p)(1-q)(1-p)$
1	0	0	1	$\frac{1}{2}$	1	1	$q(1-p)(1-q)p$
1	0	1	0	1	1	1	$q(1-p)q(1-p)$
1	0	1	1	1	1	2	$q(1-p)qp$
1	1	0	0	$\frac{1}{2}$	2	0	$qp(1-q)(1-p)$
1	1	0	1	$\frac{1}{2}$	2	1	$qp(1-q)p$
1	1	1	0	1	2	1	$qpq(1-p)$
1	1	1	1	1	2	2	$qpqp$

Table 1: Distribution of events. The sixth and seventh columns represent the treated outcomes if i lives in Milan and j in Copenhagen.

statistic for the data pair $(Y_t(1), Y_c(0))$.[21]

Thus, we proceed to compute the evaluator's conditional expectation of β_{ATE} under the assumption that the experimental result consists in the observation of $Y_t(1)$ alone. This treated outcome has three possible realisations.

First, observation $Y_t(1) = 0$ corresponds to the first four rows in Table 1. This occurs with probability

$$\Pr(Y_t(1) = 0) \;=\; (1-q)(1-p). \tag{1}$$

This observation implies that $\beta_M = 0$. There is no information about β_C, which on average is q. Thus,

$$\mathrm{E}(\beta_{ATE} | Y_t(1) = 0) \;=\; \frac{q}{2}. \tag{2}$$

[21] To verify this claim, suppose that the evaluator observes $Y_t(1) = 0$. Only the first four rows in Table 1 are possible. The evaluator then assigns conditional probability $\Pr(\beta_{ATE} = 0 | Y_t(1) = 0) = 1 - q$ to the event that $\beta_{ATE} = 0$ (first and second row of the table versus the third and fourth row). The same posterior probabilities arise when the realisation $Y_j(0)$ is also observed. Indeed, $\Pr(\beta_{ATE} = 0 | Y_t(1) = 0, Y_j(0) = 0) = 1 - q$ (see first versus third row) and $\Pr(\beta_{ATE} = 0 | Y_t(1) = 0, Y_j(0) = 1) = 1 - q$ (second versus fourth row). Analogous arguments apply when $Y_t(1) = 1$ and $Y_t(1) = 2$.

Second, $Y_t(1) = 2$ corresponds to the last four rows in Table 1. This occurs with probability

$$\Pr(Y_t(1) = 2) \; = \; qp. \tag{3}$$

The observation reveals that $\beta_M = 1$. Hence,

$$\mathrm{E}(\beta_{ATE} | Y_t(1) = 2) \; = \; \frac{1+q}{2}. \tag{4}$$

Finally, $Y_t(1) = 1$ occurs with remaining probability

$$\Pr(Y_t(1) = 1) \; = \; (1-q)\, p + q\,(1-p)\,. \tag{5}$$

Inspecting the middle eight rows of Table 1, we see that

$$\mathrm{E}(\beta_{ATE} | Y_t(1) = 1) \; = \; \frac{\frac{q(1-p)}{(1-q)p+q(1-p)} + q}{2}. \tag{6}$$

We now illustrate the welfare effects of the experiment, focusing for simplicity on the symmetric case where all values of treatment effect and baseline outcome are equally likely: $q = p = 1/2$. The qualitative conclusions we draw below (as well as most of those we derive in later sections) do not depend on the specific values of q and p that we use for illustration.

Recall that the CDF of $\mathrm{E}(\beta_{ATE} | Y_t(1))$ allows us to read the payoffs of the two parties. From equations (2), (6) and (4) we see that if $q = p = 1/2$ then, depending on the realised value of $Y_t(1)$, the conditional expectation of β_{ATE} can take values $1/4$, $1/2$ or $3/4$, with respective probabilities $1/4$, $1/2$ and $1/4$, as specified in (1), (5) and (3). The corresponding CDF is represented by the dotted curve in the middle diagram of Figure 1 below. The thin solid line in the same diagram represents the CDF of the expected average treatment effect prior to the experiment: without an experiment, with probability one, the expectation of β_{ATE} is the prior expectation, namely $q = 1/2$.

The experiment determines a clockwise rotation of the CDF of the conditional expectation of the average treatment effect, and thus a second-order stochastic dominance shift. Therefore, the evaluator's information increases. The two CDFs can be used to derive the effect of the experiment, for each possible value of the cost parameter k, on the payoffs of evaluator and researcher. This is represented in the top and bottom diagrams of Figure 1, where ΔV_R and ΔV_E denote the expected utility gain from the experiment (compared to no experiment) to researcher and evaluator, respectively. To understand the construction, recall that for a given CDF, the researcher's expected payoff is $1 - F(k)$ and the evaluator's expected payoff is $\int_k^1 (1 - F(\pi))\, d\pi$. Thus, the top diagram in Figure 1 records, for each value of the acceptance cost k, the corresponding difference between the two CDFs in the middle diagram, which is what the researcher gains from the experiment. The bottom diagram shows the area between the two CDFs, which measures the evaluator's gain. The comparison between the situations with and without the experiment illustrates the obvious fact that the evaluator can only benefit from running a non-manipulated experiment, compared to no experiment.

If the evaluator is relatively demanding ($1/2 < k < 3/4$), there would be no acceptance without the experiment. Both the evaluator and researcher then gain from the fact that the evaluator is willing

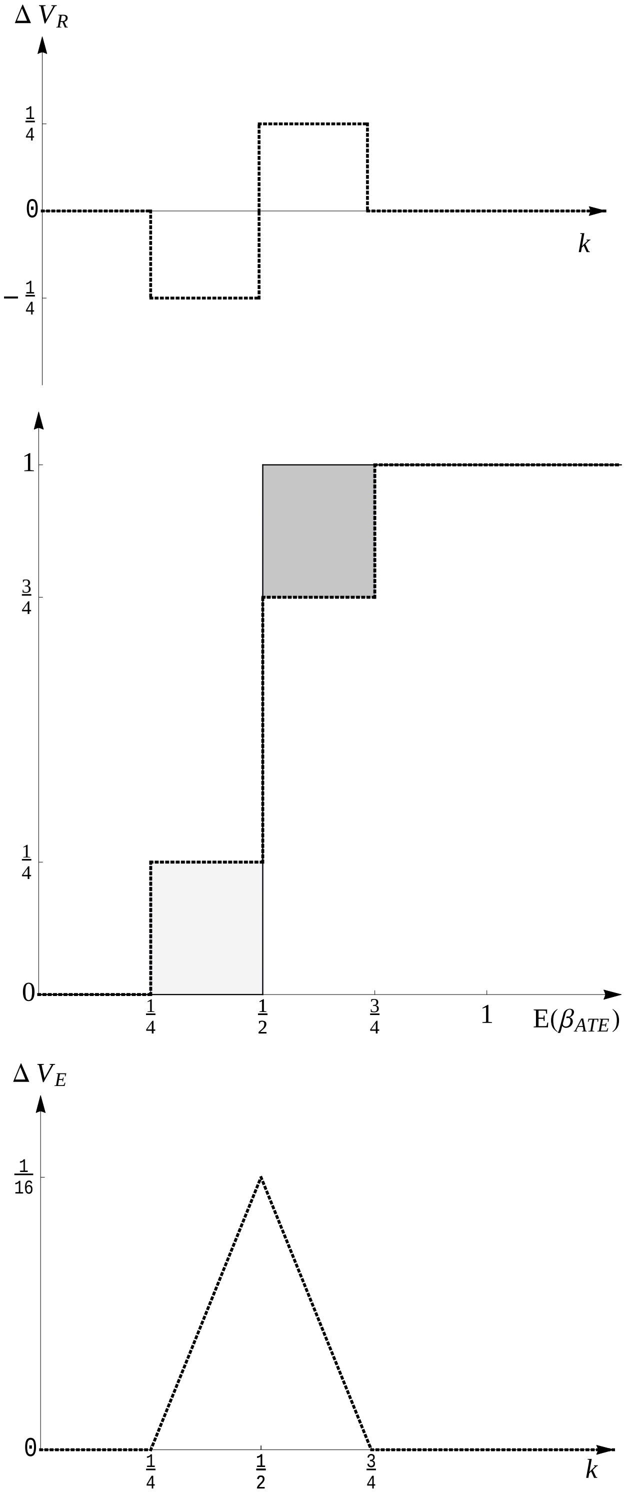

Figure 1: Information and payoff gains in the non-manipulated experiment.

to accept after the favourable evidence $Y_t(1) = 2$. On the other hand, a less demanding evaluator would accept at the prior ($1/4 < k < 1/2$). The experiment improves the evaluator's decision, but harms the researcher when the evaluator is unwilling to accept at the unfavourable evidence $Y_t(1) = 0$.

4 Selective Sampling

As mentioned before, it is private information that grants the researcher an opportunity to bias the experiment. In this section we analyse the case where the researcher has prior information about the treatment effect in one of the two sites,[22] and can pick the experimental location based on this information. This poses a challenge to the *external validity* of the experiment, as the sampling of individuals for the experiment is not random, but rather correlated with their treatment effect.

Setup. The timing is as follows:

1. The researcher privately obtains information which reveals the treatment effect in one of the two sites.
2. The researcher privately selects a site where to conduct the experiment. A random individual i from that site is assigned to treatment, $t = i$. If the experiment requires a control group, another random individual j from the site is untreated, $c = j$.[23]
3. The evaluator observes the experimental result, that is, outcome $Y_t(1)$ in an uncontrolled experiment, or both $Y_t(1)$ and $Y_c(0)$ if the experiment has a control. Based on the experimental result, the evaluator estimates the average treatment effect and chooses to accept or reject.

The evaluator knows that the researcher has access to this information, and considers the two sites equally likely. Thus, the evaluator is unable to infer any information from observing the identity of the site in which the experiment is carried out. As already mentioned, the challenge to the external validity of the experiment arises because this is held, as we argue below, in a site where individuals have an above-average treatment effect—the sampling is biased. The internal validity of the experiment is instead uncompromised, because the researcher is unable to non-randomly assign individuals to treatment and control.

We construct an equilibrium where the researcher uses the following strategy. Let $A \in \{M, C\}$ denote the site for which the researcher has observed the treatment effect. When $\beta_A = 1$, the researcher conducts the experiment in site A, while if $\beta_A = 0$, the researcher conducts it in the other

[22] This information could be obtained with an informal pilot experiment run in one of the sites.

[23] Our conclusions remain unchanged if we allow (or force) the researcher to choose treatment and control from different sites. In principle, the two individuals randomly drawn could have different realisations of their baseline types. However, given that the researcher has no private information about such realisation and the baseline types are i.i.d., the distribution of observed outcomes (and thus the inference made on its basis) is unaffected.

site. We verify below that this is indeed the optimal strategy for the researcher, given how the evaluator reacts to it. Thus, the joint behaviour we describe constitutes an equilibrium of the game.

We write $Y_t^\beta(1)$ for the treated outcome of the selected individual. Thus,

$$Y_t^\beta(1) = Y_i(0) + \max\{\beta_M, \beta_C\}. \tag{7}$$

Namely, if $\beta_A = 1$ has been observed, then the treated individual has treatment effect 1. If $\beta_A = 0$, the experiment is conducted in the other site, and the treatment effect of the other site is relevant — this effect is never less than zero.[24]

The control outcome would be $Y_j(0)$. But once again, the control is a superfluous addition to the experiment. The researcher's private selection is independent of baseline outcomes, and hence it is still the case that $Y_j(0)$ carries no useful information about β_{ATE}. Thus, we again limit ourselves to the case where the experimental result is just the outcome on the treatment group, namely $Y_t^\beta(1)$.

Inference. Again, there are three possible outcomes of the experiment.

First, consider the unfavourable outcome $Y_t^\beta(1) = 0$. This reveals from (7) that $Y_i(0) = \beta_M = \beta_C = 0$, corresponding to the first two rows in Table 1. The outcome thus occurs with probability

$$\Pr\left(Y_t^\beta(1) = 0\right) \;=\; (1-p)(1-q)^2, \tag{8}$$

and the corresponding posterior belief of the evaluator is

$$\mathrm{E}\left(\beta_{ATE}|Y_t^\beta(1) = 0\right) \;=\; 0. \tag{9}$$

This least favourable observation is thus less likely than in the non-manipulated case, but also stronger evidence against β_{ATE}. Given the strategy of the researcher, observing $Y_t^\beta(1) = 0$ leads to the conclusion that both β_A and β_B are zero.

Second, consider the favourable outcome $Y_t^\beta(1)=2$. This requires $Y_i(0)=1$ and $\max\{\beta_M, \beta_C\} = 1$, corresponding to rows 7–8 and 13–16 in Table 1. The probability is thus

$$\Pr\left(Y_t^\beta(1) = 2\right) \;=\; pq(2-q), \tag{10}$$

and inspection of the rows in the table helps to establish that

$$\mathrm{E}\left(\beta_{ATE}|Y_t^\beta(1) = 2\right) \;=\; \frac{(2-2q)\,\frac{1}{2} + q}{2-q} \;=\; \frac{1}{2-q}. \tag{11}$$

Since $\beta_M = 1$ implies $\max\{\beta_M, \beta_C\} = 1$, this favourable outcome is of course more likely than under no manipulation. Comparison of (11) with (4) shows that the posterior mean is less extreme, as $0 < q < 1$ implies that $\mathrm{E}\left(\beta_{ATE}|Y_t^\beta(1) = 2\right)$ is strictly smaller than $\mathrm{E}\left(\beta_{ATE}|Y_t(1) = 2\right)$.

[24] Observe that we could have alternatively assumed that the researcher privately obtains information about the treatment effect in both sites, and then selects to conduct the experiment in a site where the treatment effect is maximal.

The rational evaluator discounts the evidence because the researcher's biased selection strategy is anticipated. In fact, for some parameter values, $\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=2\right)$ is lower than the middling $\mathrm{E}\left(\beta_{ATE}|Y_t(1)=1\right)$ computed earlier under no manipulation.

Finally, the outcome $Y_t^{\beta}(1)=1$ occurs with the remaining probability,

$$\Pr\left(Y_t^{\beta}(1)=1\right) \;=\; (1-p)\,q\,(2-q)+p\,(1-q)^2\,. \tag{12}$$

Inspection of the remaining eight rows in the table provides

$$\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=1\right) \;=\; \frac{(1-p)\,q\,(2-2q)\,\frac{1}{2}+(1-p)\,q^2}{(1-p)\,q\,(2-q)+p\,(1-q)^2}. \tag{13}$$

The comparison of this case with the corresponding case under no manipulation, $Y_t(1)=1$, is not straightforward. First, $\Pr\left(Y_t^{\beta}(1)=1\right)=\Pr\left(Y_t(1)=1\right)+(1-2p)(1-q)\,q$, so that observing the intermediate outcome can be more or less likely than under no manipulation, according to whether $p<1/2$ or $p>1/2$. Second, and more importantly, the comparisons between (13) and (6) and between (13) and (2) are far from immediate: the conditional expectation $\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=1\right)$ is not always smaller than the conditional expectation associated to the intermediate outcome in the non-manipulated case, $\mathrm{E}\left(\beta_{ATE}|Y_t(1)=1\right)$, nor always greater than the conditional expectation associated to the least favourable outcome in the non-manipulated case, $\mathrm{E}\left(\beta_{ATE}|Y_t(1)=0\right)$. This makes the comparison between the case of no manipulation and the case of selective sampling a subtle one, as we discuss below.

Our calculations show that under selective sampling the evaluator may raise or lower (depending on the values of p and q) the standard of acceptance, compared to the non-manipulated case. However, for every k, the evaluator's strategy is again monotone in the observed outcome: if acceptance occurs at a value of $Y_t^{\beta}(1)$ then it also occurs at larger values of $Y_t^{\beta}(1)$. Thus, it is optimal for the researcher to treat individual i when the treatment effect in site A is expected to be larger than that in site B, that is, if and only if $\beta_A=1$. This verifies that we have characterised an equilibrium of the game.

Welfare. A general consideration allows us to analyse the impact of selective sampling on the evaluator's payoff. We have so far derived two evaluator strategies, when seen as whether to accept after observing a given outcome. The "trusting" strategy was derived in the previous section as the optimal reaction to a non-manipulated experiment. The "untrusting" strategy has been derived in the present section in reaction to the researcher's strategic site selection.

Now, consider the event that the outcome of the treated individual under manipulation, $Y_t^{\beta}(1)$, is different from the non-manipulated outcome, $Y_t(1)$ (corresponding to rows $3-4$ and $7-8$ in Table 1). This is the event where selection affects the evaluator's payoff, assuming the evaluator's strategy is held fixed to either the trusting or untrusting strategy. Inspection of the relevant rows in Table 1 shows that $Y_t^{\beta}(1)>Y_t(1)$ and $\beta_{ATE}=1/2$ in this event. Thus, manipulation can only lead to additional acceptances, and if so to a conditional evaluator gain of $(1/2) - k$. By implication, the trusting evaluator can only benefit from manipulation when $k < 1/2$, and the untrusting evaluator can only be harmed by manipulation when $k > 1/2$. We conclude that a fortiori the rational evaluator, who optimally reacts to the researcher's strategy, can also only benefit from manipulation for $k < 1/2$, and only be harmed by it for $k > 1/2$.[25]

The impact of manipulation on the researcher's payoff is more subtle. The fact that, compared to the case of no manipulation, the least and most favourable outcomes both lead to more pessimistic inferences, and are less and more likely, respectively, immediately implies that for extreme values of k the researcher is always (for all values of q and p) harmed by the opportunity of selective sampling. However, the exact bounds on k defining its "extreme values" depend on the specific values of q and p. Moreover, while for a large region of values of q and p the researcher benefits from manipulation for a connected set of intermediate values of k, there is a also a region of values of q and p for which this set is not connected, so that the researcher is also harmed for some intermediate values of k. We report precise statements in the appendix. In the next result, we limit ourselves to recording the impact on the researcher's payoff for the symmetric case $p = \frac{1}{2}$.

Proposition 1 (Welfare impact of selective sampling) *Compared to the non-manipulated experiment, selective sampling harms or benefits the evaluator according to whether k is larger or smaller than*

$$\hat{k}^{\beta} = \frac{1}{2}.$$

Moreover, if $p = 1/2$, selective sampling harms or benefits the researcher according to whether k is outside or inside the interval

$$\left[\mathrm{E}(\beta_{ATE}|Y_t(1) = 0), \mathrm{E}(\beta_{ATE}|Y_t^{\beta}(1) = 2)\right].$$

We illustrate the result in Figure 2, assuming $q = p = 1/2$. In this figure, ΔV_R and ΔV_E denote the researcher's and evaluator's expected utility gains from playing the equilibrium of the game with a strategically selected site instead of the equilibrium of the non-manipulated experiment. Since $q = p = 1/2$, the posterior expectation following the intermediate outcome is unchanged by manipulation, that is, we have $\mathrm{E}(\beta_{ATE}|Y_t^{\beta}(1) = 1) = \mathrm{E}(\beta_{ATE}|Y_t(1) = 1)$. This implies that the CDF corresponding to the manipulated case (solid CDF in the middle diagram) crosses the CDF corresponding to no manipulation (dotted CDF) only twice. Consistent with the proposition, the top and bottom diagrams illustrate that the researcher is harmed by manipulation for extreme values of the cost parameter k, while the evaluator benefits only for $k < 1/2$.

The maximum benefit for the evaluator obtains for a low level of the acceptance cost, $k = \mathrm{E}(\beta_{ATE}|Y_t(1) = 0)$. Manipulation helps this less demanding evaluator because it makes the unfavourable evidence more compelling. As noted before, the outcome $Y_t^{\beta}(1) = 0$ comes to reveal that not only $\beta_M = 0$ but also $\beta_C = 0$, once selective sampling is possible. As a result, around the

[25] For some values of the parameters p, q, k, the evaluator accepts in the same rows of Table 1, and is thus unaffected by the research bias.

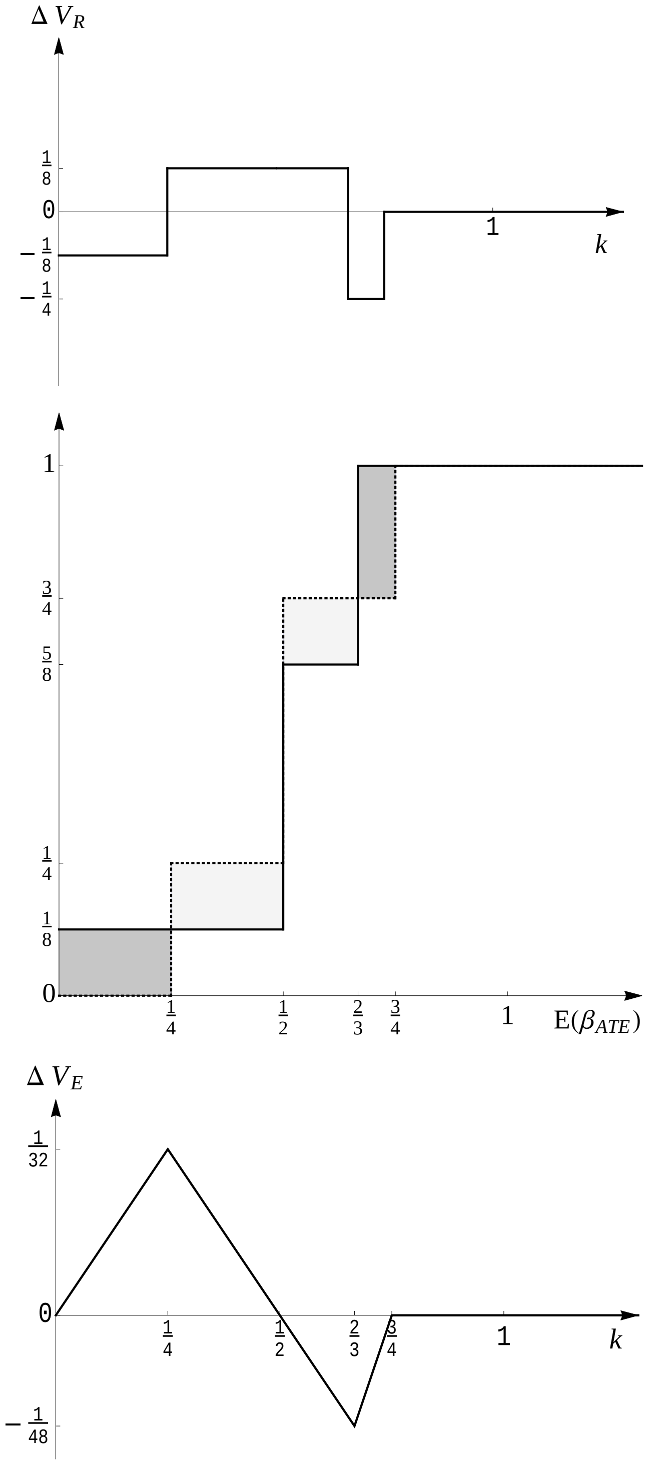

Figure 2: Impact of selective sampling in the symmetric case.

point where $k = \mathrm{E}\left(\beta_{ATE} | Y_t\left(1\right) = 0\right)$, the solid CDF is a clockwise rotation of the dotted CDF—manipulation increases information “locally” around that point.

However, the opposite happens for the rather high acceptance cost $k = \mathrm{E}(\beta_{ATE} | Y_t^{\beta}(1) = 2)$. The bias allows the researcher to attain more often the optimal outcome $Y_t^{\beta}\left(1\right) = 2$, but for the same reason the outcome is contaminated as evidence that β_{ATE} should be high. Information is lost “at the top,” as here the dotted CDF is a clockwise rotation of the solid CDF.

Finally, the model predicts that selective sampling creates a *credibility crisis* when the evaluator is demanding (high *k*). In Figure 2, when $k \in \left(2/3, 3/4\right)$, both the researcher and the evaluator are harmed by the opportunity to manipulate the experiment. They both have private motivations to commit the researcher to avoid obtaining advance information about the treatment effect. It is but difficult. Once the evaluator “trusts” the researcher and evaluates the evidence as if from a non-manipulated experiment, the researcher is tempted to perform the manipulation because this raises the probability to get the favourable outcome $Y_t^{\beta}\left(1\right) = 2$, which is necessary to convince a demanding evaluator.

5 Selective Assignment

We turn now to the second main form of research manipulation, based on private information about the baseline outcomes. If the researcher can pick the individual to assign to treatment based on this information, a challenge to the *internal validity* of the experiment arises, as treatment status becomes correlated with baseline outcome.

Setup. The timing in this model is as follows:

1. The researcher privately observes the baseline outcome of one individual in one site.[26]
2. The researcher selects a site and an individual from this site for treatment. If the experiment requires a control group, the researcher privately selects from the same site also another individual who does not receive the treatment.
3. The evaluator observes the experimental result, that is, outcome $Y_t\left(1\right)$ in an uncontrolled experiment, or both $Y_t\left(1\right)$ and $Y_c\left(0\right)$ if the experiment has a control. Based on the experimental result, the evaluator estimates the average treatment effect and chooses to accept or reject.

The researcher learns nothing about treatment effects, but obtains (perfect) information concerning the baseline outcome of a random individual *i*. The researcher can then assign individual *i* or

[26] This assumption is an extreme version of the idea that the researcher can predict the baseline health on the basis of observed correlates. This simplifying assumption should not be taken literally, as it is a premise of the RCT that one cannot observe both the untreated and treated outcomes for the same subject.

a random individual j to treatment (or control, if a control must be provided), based on whether $Y_i(0) = 0$ or $Y_i(0) = 1$. As already mentioned, this undermines the internal validity of the experiment, as randomisation of the assignment to treatment (and control, if a control is required) is compromised—the treated individual has an above-average baseline outcome.

5.1 Uncontrolled Experiment

In this section we focus exclusively on the simpler non-controlled experiment. The case with a control group is analysed in Section 5.2. Without loss of generality, we assume that the experiment is carried out in site M, i.e. that both i and j reside in M.[27]

Like in the case of selective sampling, we naturally guess that the researcher aims at getting as large an experimental result as possible, and thus treats individual $t = i$ when $Y_i(0) = 1$ and $t = j$ when $Y_i(0) = 0$. Here, too, this will turn out to be the researcher's optimal strategy, given the evaluator's best response to it. We denote by

$$Y_t^{\varepsilon}(1) = \max\left\{Y_i(0), Y_j(0)\right\} + \beta_M \tag{14}$$

the outcome of the treated individual.

Note, here and in the non-manipulated experiment, the evaluator sees an outcome which is β_M plus some noise. In the non-manipulated experiment, the noise is $Y_i(0)$, while here it is the systematically higher $\max\left\{Y_i(0), Y_j(0)\right\}$.

Inference. There are three possible outcomes of the experiment.

First consider the unfavourable outcome $Y_t^{\varepsilon}(1) = 0$. This requires that $Y_i(0) = Y_j(0) = 0$ and $\beta_M = 0$, corresponding to rows 1 and 3 in Table 1. It occurs with probability

$$\Pr\left(Y_t^{\varepsilon}(1) = 0\right) \;=\; (1-q)(1-p)^2. \tag{15}$$

Since $Y_j(0) = 0$ is an extra requirement, we easily see that $\Pr\left(Y_t^{\varepsilon}(1) = 0\right) < \Pr\left(Y_t(1) = 0\right)$. The outcome reveals that the treatment effect on the treated individual is zero. Thus, the conditional expectation of β_{ATE} is the same as in the non-manipulation benchmark:

$$\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 0\right) \;=\; \frac{q}{2}. \tag{16}$$

Next, consider the favourable outcome $Y_t^{\varepsilon}(1) = 2$. This corresponds to rows 10 and 12–16 in Table 1. The outcome occurs with probability

$$\Pr\left(Y_t^{\varepsilon}(1) = 2\right) \;=\; qp(2-p). \tag{17}$$

[27] In principle, the two individuals randomly drawn could belong to different sites. However, given that the researcher has no private information about the treatment effect in the site in which j resides, the distribution of observed outcomes (and hence the evaluator's inference made on its basis) is unaffected.

Since $Y_i(0) = 1$ implies $\max\{Y_i(0), Y_j(0)\} = 1$, the favourable outcome is more likely than under no manipulation: $\Pr(Y_t^{\varepsilon}(1) = 2) > \Pr(Y_t(1) = 2)$. The outcome reveals that $\beta_M = 1$ and carries no information about β_C. Thus, also here the conditional expectation of β_{ATE} is the same as in the non-manipulation benchmark:

$$\mathrm{E}(\beta_{ATE}|Y_t^{\varepsilon}(1) = 2) \; = \; \frac{1+q}{2}. \tag{18}$$

Finally, the outcome $Y_t^{\varepsilon}(1) = 1$ occurs with the remaining probability,

$$\Pr(Y_t^{\varepsilon}(1) = 1) \; = \; (1-q)p(2-p) + q(1-p)^2. \tag{19}$$

Simple algebra using (19) and (5) shows that $\Pr(Y_t^{\varepsilon}(1) = 1) = \Pr(Y_t(1) = 1) + (1-2q)\,p\,(1-p)$. Thus, the probability of the intermediate outcome is more or less than under no manipulation, according to whether $q < 1/2$ or $q > 1/2$, respectively. Using rows 2, 4–9, and 11 in Table 1, we see that the posterior expectation on the average treatment effect is

$$\mathrm{E}(\beta_{ATE}|Y_t^{\varepsilon}(1) = 1) \; = \; \frac{(1-q)q\frac{1}{2} + q^2(1-p)^2}{(1-q)p(2-p) + q(1-p)^2}. \tag{20}$$

By comparing expressions (6) and (20) one can verify that the latter is smaller, for all values of q and p. Under manipulation, the intermediate outcome leads to a more pessimistic inference than under no manipulation. This was to be anticipated, since the higher noise realisation $\max\{Y_i(0), Y_j(0)\} \geq Y_i(0)$ makes the outcome $Y_t^{\varepsilon}(1) = 1$ more frequent when $\beta_M = 0$ and less frequent when $\beta_M = 1$.

Unlike in the case of selective sampling, the analysis shows that under selective assignment the evaluator always (for every p and q) raises the standard of acceptance. But for every k, as before, the evaluator's strategy is still monotone in the observed outcome: if acceptance occurs at a value of $Y_t^{\varepsilon}(1)$ then it also occurs at larger values of $Y_t^{\varepsilon}(1)$. This implies that we have correctly identified an equilibrium of the game: it is indeed optimal for the researcher to treat individual i when the individual's baseline outcome is high, $Y_i(0) = 1$.

Welfare. The analysis above has straightforward implications for the evaluator's and researcher's equilibrium payoffs. In the manipulated experiment, the least and most favourable experimental results, $Y_t^{\varepsilon}(1) = 0$ and $Y_t^{\varepsilon}(1) = 2$, are respectively less and more likely than in the non-manipulated experiment of the previous section, but yield the same posterior expected average treatment effect. Moreover, the intermediate result, $Y_t^{\varepsilon}(1) = 1$, results in a less favourable posterior. This implies that the CDF of the posterior expectation of β_{ATE} under manipulation crosses the CDF of the non-manipulated case twice, first from below at $\mathrm{E}(\beta_{ATE}|Y_t^{\varepsilon}(1) = 1)$, and then from above at $\mathrm{E}(\beta_{ATE}|Y_t(1) = 1)$.

The above facts yield the following conclusion. The evaluator is worse off than under no manipulation for low k, and better off for high k. Moreover, the researcher is better off for extreme values of k, and worse off for intermediate values of k. In other words, the welfare conclusions drawn from the analysis of selective sampling are reversed, when we consider selective assignment instead. As we discuss below, this is because in this case, contrary to the case of selective sampling, manipulation locally increases information “at the top” rather than “at the bottom.”

Proposition 2 (Welfare impact of selective assignment) *Compared to the non-manipulated experiment, selective assignment harms or benefits the researcher according to whether k is inside or outside the interval*

$$\left[\mathrm{E}(\beta_{ATE}|Y_t^{\varepsilon}(1)=1), \mathrm{E}(\beta_{ATE}|Y_t(1)=1)\right].$$

Moreover, for a value inside this interval,

$$\hat{k}^{\varepsilon} = \frac{(1-q)qp + q\,(1+q)\,(1-p)^2}{(1-q)p + q(1-p)^2},$$

selective assignment harms or benefits the evaluator according to whether k is smaller or larger than $\hat{k}^{\varepsilon}$.

We illustrate the conclusions of the proposition in Figure 3 for the symmetric case $q = p = 1/2$. In this case, equations (15)–(18) show that the conditional expectation of β_{ATE} can take values $1/4$, $3/8$ or $3/4$, with respective probabilities $1/8$, $1/2$ and $3/8$. In the middle diagram of Figure 3 we contrast the CDF of the expected average treatment effect under selective assignment (solid curve) to the CDF corresponding to the non-manipulated experiment (dotted curve). The top and bottom diagrams, where ΔV_R and ΔV_E denote the expected utility gains from playing the equilibrium of the game with selective assignment instead of the equilibrium of the non-manipulated experiment, illustrate the fact that the researcher benefits from manipulation (only) for extreme values of the cost parameter k, while the evaluator benefits (only) for sufficiently large values of k.

The maximum benefit for the evaluator obtains at the middling $k = \mathrm{E}\left(\beta_{ATE}|Y_t\left(1\right)=1\right) = 1/2$. With this k, the evaluator would be indifferent between accepting and rejecting after observing the non-manipulated intermediate outcome $Y_t\left(1\right) = 1$. Manipulation helps the evaluator break this indifference by indirectly providing information about the counterfactual baseline outcome of the treated individual: as is apparent from the middle diagram in Figure 3, around the point where $k = 1/2$ the solid CDF is a clockwise rotation of the dotted CDF—manipulation increases information locally around that point. The opposite happens around the point $k = \mathrm{E}\left(\beta_{ATE}|Y_t^{\varepsilon}\left(1\right)=1\right) = 3/8$. Here, manipulation locally decreases information—the dotted CDF is a clockwise rotation of the solid CDF.

Intuitively, what helps a demanding evaluator to make better decisions in this case is that the desirable state where $\beta_M = 1$ is easier to detect. Upward manipulation of the baseline outcome implies that this state results less often in the intermediate treated outcome of 1. The researcher’s bias affects the distribution of the noise term—the baseline outcome—in such a way that $\beta_M = 1$ more often distinguishes itself in the high treated outcome 2.

Observe that also with this type of manipulation there can be a scope for a *credibility crisis* where both parties are harmed. In the example of Figure 3, this arises when $k \in (3/8, 9/24)$.

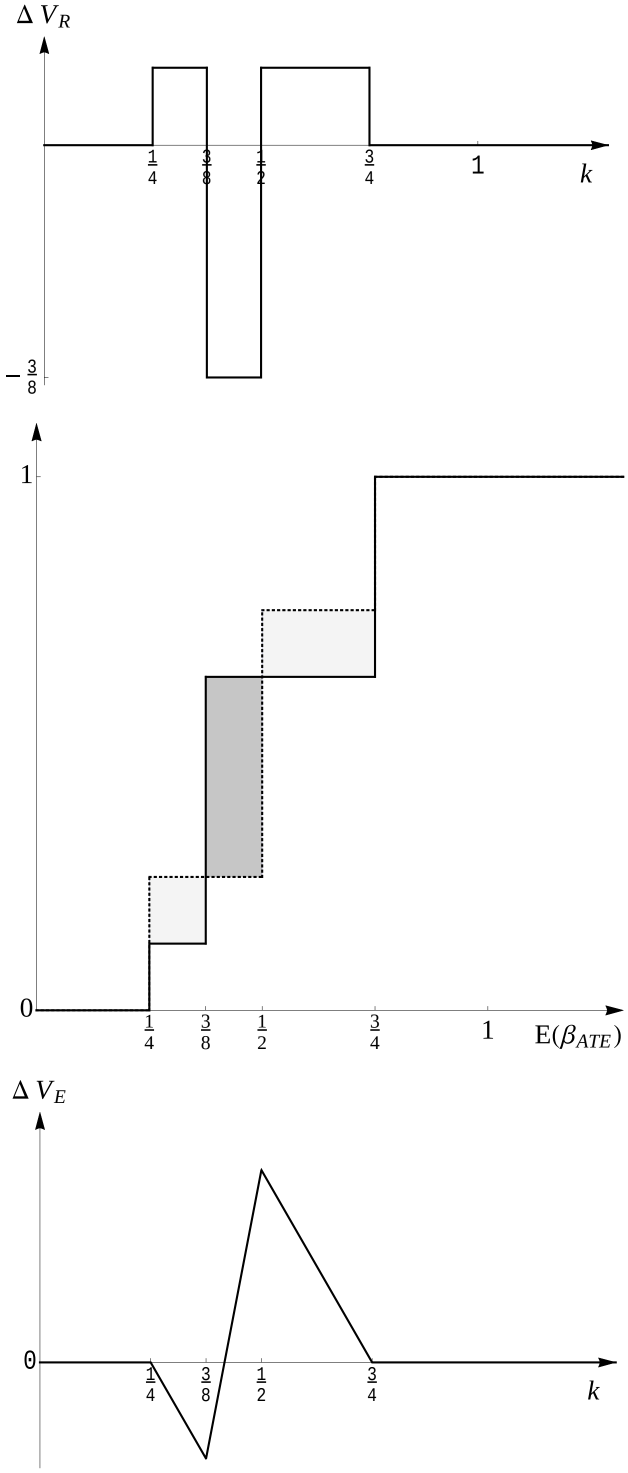

Figure 3: Impact of selective assignment in the symmetric case.

5.2 Value of Control

As we have seen above, given that in our model the distribution of baseline outcomes is known, adding a control group to the experiment does not benefit the evaluator if the experiment is not manipulated, or if manipulation only occurs via selective sampling. However, as we now show, in the presence of selective assignment the evaluator fares better with a controlled experiment, that is, if in addition to the outcome of the treated individual the evaluator is able to observe the baseline outcome of an untreated individual.

As assumed earlier, the experiment takes place in site M, with the researcher observing $Y_i(0)$. Again, we suppose that the researcher selects to treat individual $t = i$ if $Y_i(0) = 1$ and $t = j$ otherwise. As before, we denote the outcome of the treated individual by $Y_t^{\varepsilon}(1) = \max\{Y_i(0), Y_j(0)\} + \beta_M$.

However, here we assume that the experiment must have a control group, too. Thus, the researcher must also select from site M an individual c not to receive the treatment. We suppose that the researcher chooses $c = i$ when $Y_i(0) = 0$ and $c = j$ when $Y_i(0) = 1$. The baseline outcome of individual c is therefore

$$Y_c^{\varepsilon}(0) = \min\left\{Y_i(0), Y_j(0)\right\}.$$

Notice that the evaluator's posterior inferences in the cases $Y_t^{\varepsilon}(1) = 0$ and $Y_t^{\varepsilon}(1) = 2$ are identical to those in the corresponding cases without a control group. Here, too, these observations perfectly reveal the treatment effect β_M, namely $\beta_M = 0$ in the first case and $\beta_M = 1$ in the second, and there is no further information in the baseline outcome $Y_c^{\varepsilon}(0)$ of the untreated individual.

However, observing $Y_c^{\varepsilon}(0)$ besides $Y_t^{\varepsilon}(1)$ improves the evaluator's inference whenever $Y_t^{\varepsilon}(1) = 1$. Indeed, when observing $Y_t^{\varepsilon}(1) = 1$ and $Y_c^{\varepsilon}(0) = 1$, the evaluator learns that the treatment effect on the treated is zero.[28] On the flip side, observing $Y_t^{\varepsilon}(1) = 1$ and $Y_c^{\varepsilon}(0) = 0$ makes the evaluator more optimistic (the conditional expectation of β_{ATE} is higher) than observing $Y_t^{\varepsilon}(1) = 1$ alone.

The monotonicity of the evaluator's posterior belief about β_{ATE} again verifies that the suggested selection strategy of the researcher is indeed part of an equilibrium.

We now turn to the construction of the CDF of the evaluator's posterior expectation of β_{ATE}. Although there are six possible outcome realisations of the pair $(Y_t^{\varepsilon}(1), Y_c^{\varepsilon}(0))$, the evaluator's posterior expectation is again concentrated on three possible values.

First, observations with $Y_t^{\varepsilon}(1) = 0$ or $(Y_t^{\varepsilon}(1), Y_c^{\varepsilon}(0)) = (1, 1)$ reveal that $\beta_M = 0$. This happens with probability

$$\Pr\left(Y_t^{\varepsilon}(1) = 0 \text{ or } Y_t^{\varepsilon}(1) = Y_c^{\varepsilon}(0) = 1\right) \;=\; (1-q)(1-p)^2 + (1-q)\,p^2, \tag{21}$$

and the evaluator's posterior is again

$$\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 0\right) \;=\; \mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = Y_c^{\varepsilon}(0) = 1\right) \;=\; \frac{q}{2}. \tag{22}$$

[28] Namely, $Y_c^{\varepsilon}(0) = 1$ implies that $Y_t^{\varepsilon}(0) = \max\left\{Y_i(0), Y_j(0)\right\} \geq \min\left\{Y_i(0), Y_j(0)\right\} = 1$.

Note, comparing (1) and (21), that the probability of this can be larger or smaller than under no manipulation, according to whether $p < 1/2$ or $p > 1/2$, respectively. We return to this point below.

Next, observations with $Y_t^{\varepsilon}(1) = 2$ reveal that $\beta_M = 1$. The probability is again

$$\Pr(Y_t^{\varepsilon}(1) = 2) \;=\; qp(2-p), \tag{23}$$

and the posterior is again

$$\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 2\right) \;=\; \frac{1+q}{2}. \tag{24}$$

Finally, the remaining observation that $(Y_t^{\varepsilon}(1), Y_c^{\varepsilon}(0)) = (1,0)$ arises either because $\beta_M = 0$ and $Y_i(0) \neq Y_j(0)$, or because $\beta_M = 1$ and $Y_i(0) = Y_j(0) = 0$. This corresponds to rows 2, 4–5, 7, 9 and 11 in Table 1. The probability is

$$\Pr(Y_t^{\varepsilon}(1) = 1, Y_c^{\varepsilon}(0) = 0) \;=\; 2(1-q)\,p(1-p) + q(1-p)^2, \tag{25}$$

and the posterior expectation is

$$\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 1, Y_c^{\varepsilon}(0) = 0\right) \;=\; \frac{(1-q)q(1+p)\frac{1}{2} + q^2(1-p)}{2(1-q)\,p + q(1-p)}. \tag{26}$$

Comparing (26) with (6) and (20) we note that the expectation $\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 1, Y_c^{\varepsilon}(0) = 0\right)$ is smaller than $\mathrm{E}\left(\beta_{ATE} | Y_t(1) = 1\right)$ but larger than $\mathrm{E}\left(\beta_{ATE} | Y_t^{\varepsilon}(1) = 1\right)$. Thus, despite an observed difference of one (the maximum value of β_{ATE}) between the outcomes of the treated and the untreated individual, the evaluator draws a more pessimistic conclusion about the average treatment effect than under no manipulation, just as in the manipulated case without a control group. However, the inference is less pessimistic than in the latter case, because observing $Y_t^{\varepsilon}(1) = 1$ and $Y_c^{\varepsilon}(0) = 0$ rather than $Y_t^{\varepsilon}(1) = 1$ alone rules out one of the possibilities (namely $Y_t^{\varepsilon}(1) = 1$ and $Y_c^{\varepsilon}(0) = 1$) under which $\beta_{ATE} = 0$.

The bottom line of the analysis is simple. Even though a controlled experiment gives no extra benefit to the evaluator under no manipulation, in the presence of selective assignment adding a control improves the evaluator's welfare. The addition of the control allows the evaluator to obtain better knowledge of the (counterfactual) baseline outcome of the treated individual.

Proposition 3 (Value of controlled experiment under selective assignment) *Under selective assignment, compared to the uncontrolled experiment, the controlled experiment weakly benefits the evaluator for every k, and strictly benefits the evaluator when k is in the interval*

$$\left[\mathrm{E}(\beta_{ATE}|Y_t(1) = 0), \mathrm{E}(\beta_{ATE}|Y_t^{\varepsilon}(1) = 1, Y_c^{\varepsilon}(0) = 0)\right].$$

We illustrate our conclusions in Figure 4, where assuming $q = 1/2$ and either $p = 3/8$ (left) or $p = 5/8$ (right) we contrast the information structure of the previous section (solid CDFs in top diagrams) with those under no manipulation (dotted CDFs in top diagrams) and in the controlled

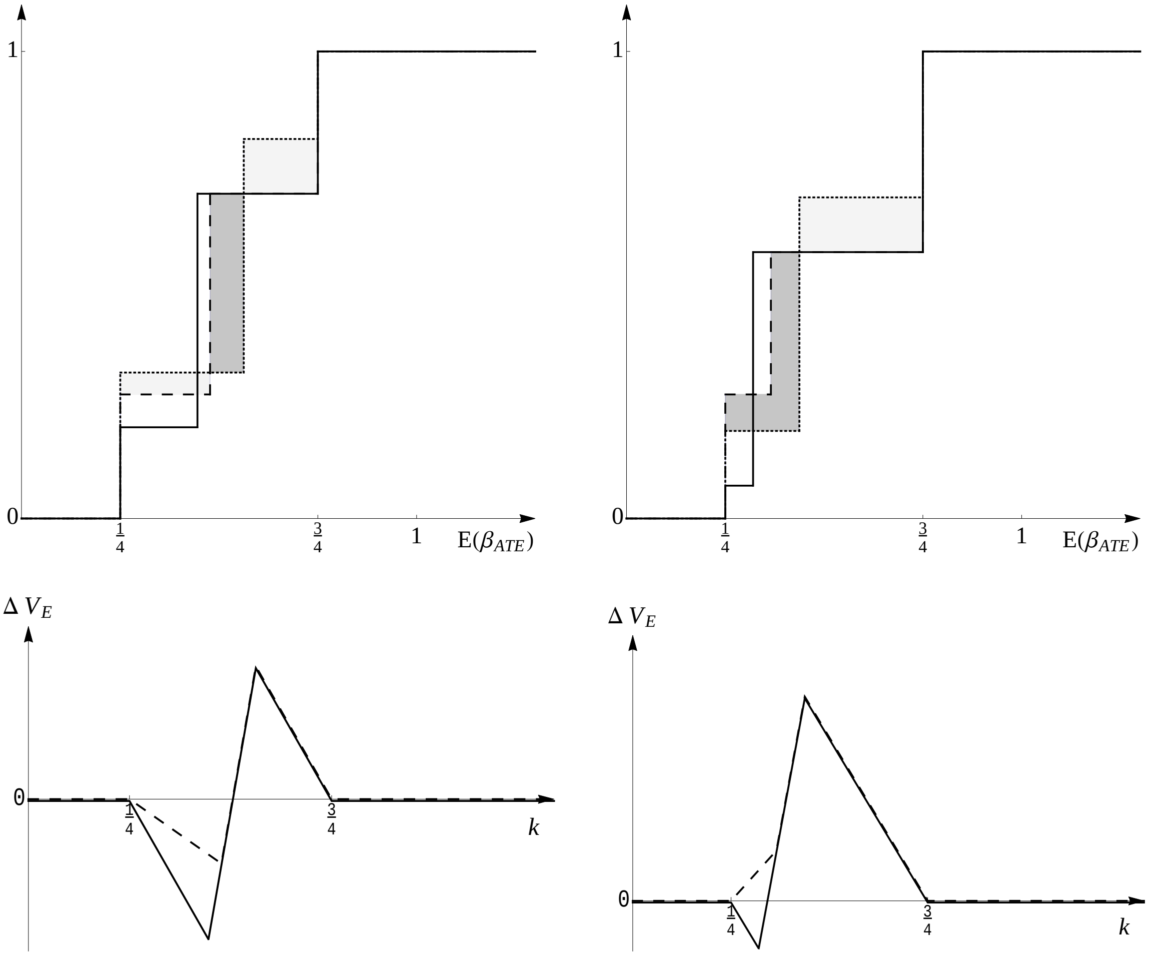

Figure 4: Adding control group improves inference under selective assignment.

manipulated experiment (dashed CDFs in top diagrams). The dashed CDFs are clockwise rotations of the solid CDFs around the point where $k = \mathrm{E}\left(\beta_{ATE}|Y_t^{\varepsilon}(1) = 1\right)$. That is, the controlled experiment provides better information precisely where the evaluator is indifferent (in the uncontrolled experiment) between accepting and rejecting after observing $Y_t^{\varepsilon}(1) = 1$.

The improvement in evaluator's welfare is sizeable: for $p = 1/2$, the negative effect of manipulation (compared to no manipulation) is completely undone by adding a control. For $p < 1/2$, it is partly undone, while for $p > 1/2$ it is even reversed—in this case, manipulation always weakly benefits the evaluator, and strictly benefits for intermediate values of the cost parameter, $(q/2) < k < (1+q)/2$. This can be seen in the bottom diagrams of Figure 4, where we represent the evaluator's payoff difference between uncontrolled and non-manipulated experiment (solid curve, as in the bottom diagram of Figure 3) and between controlled and non-manipulated experiment (dashed curve).

6 Selective Reporting

In this section we discuss a third kind of manipulation, whereby the researcher privately performs two experiments, one in each site, and then reports only one of the two. We limit our attention to the case of an uncontrolled experiment.

Setup. The timing is as follows:

1. The researcher privately observes the treated outcomes $Y_i(1)$ and $Y_j(1)$ for two individuals, i from Milan and j from Copenhagen.

2. The researcher selects a site t.[29]

3. The evaluator observes the outcome $Y_t(1)$ of the treatment in site t, and based on this, estimates the average treatment effect and chooses to accept or reject.

The researcher's ability to conceal one of the two experiments poses a challenge to both internal and external validity. Similar to the case of selective assignment, this post hoc outcome selection poses a challenge to the internal validity of the experiment, because the observed $Y_i(1)$ and $Y_j(1)$ carry information concerning $Y_i(0)$ and $Y_j(0)$. Moreover, similar to the case of selective sampling, the choice of location of the experiment is not independent of the treatment effect, as $Y_i(1)$ and $Y_j(1)$ also carry information about β_M and β_C. Indeed, once again we guess that the researcher aims at obtaining a large experimental result, and thus treats individual $t = i$ when $Y_i(1) > Y_j(1)$

[29] This form of hiding evidence from the un-selected site is an extreme form of selective reporting, but allows us to illustrate the effects within our model.

and individual $t = j$ otherwise. Here, too, this will turn out to be the researcher's optimal strategy, given the evaluator's best response to it. We denote by

$$Y_t^\rho(1) = \max\left\{Y_i(1), Y_j(1)\right\} \tag{27}$$

the outcome of the treated individual.

Inference. Once again there are three possible experimental results.

First consider the unfavourable result, $Y_t^\rho(1) = 0$. This requires both $Y_i(0) = Y_j(0) = 0$ and $\beta_M = \beta_C = 0$, corresponding to the first row in Table 1, and occurs with probability

$$\Pr\left(Y_t^\rho(1) = 0\right) \;=\; (1-q)^2(1-p)^2 . \tag{28}$$

Clearly, $\Pr\left(Y_t^\rho(1) = 0\right) < \Pr(Y_t(1) = 0)$, and similarly to the case of selective sampling, the result reveals that the average treatment effect is zero:

$$\mathrm{E}\left(\beta_{ATE}|Y_t^\rho(1) = 0\right) \;=\; 0. \tag{29}$$

Next, consider the favourable outcome, $Y_t^\rho(1) = 2$. This corresponds to rows 4, 8 and 12–16 in Table 1. The outcome occurs with probability

$$\Pr\left(Y_t^\rho(1) = 2\right) \;=\; qp(2-qp), \tag{30}$$

which is greater than under no manipulation: $\Pr\left(Y_t^\rho(1) = 2\right) > \Pr(Y_t(1) = 2)$. The outcome reveals that the treatment effect in the site of the experiment is 1, but provides unfavourable evidence about the treatment effect in the other site. Indeed,

$$\mathrm{E}\left(\beta_{ATE}|Y_t^\rho(1) = 2\right) \;=\; \frac{2qp(1-q)\frac{1}{2} + q^2p(2-p)}{qp(2-qp)} \;=\; \frac{1+q-qp}{2-qp}, \tag{31}$$

which is strictly smaller than $\mathrm{E}(\beta_{ATE}|Y_t(1) = 2)$, albeit always greater than $\mathrm{E}(\beta_{ATE}|Y_t(1) = 1)$.

Finally, let us consider the intermediate outcome, $Y_t^\rho(1) = 1$. This occurs with the remaining probability,

$$\Pr\left(Y_t^\rho(1) = 1\right) \;=\; 1-(1-q)^2(1-p)^2 - qp(2-qp). \tag{32}$$

Comparing (32) with (5), one can show that the probability of the intermediate outcome is more or less than under no manipulation, according to whether $q + p < 1$ or $q + p > 1$, respectively. Using the remaining rows 2–3, 5–7, and 9–11 in Table 1, we obtain the expected average treatment effect,

$$\mathrm{E}\left(\beta_{ATE}|Y_t^\rho(1) = 1\right) \;=\; \frac{2(1-q)(1-p)q\frac{1}{2} + q^2(1-p)^2}{1-(1-q)^2(1-p)^2 - qp(2-qp)}. \tag{33}$$

For all values of q and p, the latter expression is always smaller than the corresponding non-manipulated expectation in (6). As in the case of selective assignment, the intermediate outcome leads to a more pessimistic inference than under no manipulation. In fact, the inference can be even worse than under the unfavourable result under no manipulation, as comparison of (2) and (33) tells us that $\mathrm{E}\left(\beta_{\mathrm{ATE}}|Y_t^{\rho}(1)=1\right)$ can be larger or smaller than $\mathrm{E}\left(\beta_{\mathrm{ATE}}|Y_t(1)=0\right)$ depending on parameter values. More precisely, we have

$$\mathrm{E}\left(\beta_{\mathrm{ATE}}|Y_t^{\rho}(1)=1\right) < \mathrm{E}\left(\beta_{\mathrm{ATE}}|Y_t(1)=0\right)$$

if and only if both

$$p > 2-\sqrt{2} \quad \text{and} \quad q < 1-\frac{1-p}{\sqrt{2p-1}}. \tag{34}$$

As in the case of selective assignment, under selective reporting the evaluator always raises the standard of acceptance. Moreover, for every k, as before, the acceptance strategy is monotone in the observed outcome: if acceptance occurs at a value of $Y_t^{\rho}(1)$, it also occurs at larger values of $Y_t^{\rho}(1)$. Thus here, too, we have identified an equilibrium of the game.

Welfare. The welfare impact of selective reporting resembles in some respects the one of selective sampling. Here, too, the least and most favourable outcomes both lead to more pessimistic inferences, and are less and more likely, respectively. Thus, for extreme values of the cost parameter k, the welfare implications are similar. The researcher suffers from manipulation, while the evaluator is harmed by or benefits from manipulation for, respectively, high and low extreme values of k. Thus, this model also predicts a credibility crisis due to manipulation when the acceptance cost is high. Again, the exact bounds defining "high" depend on the specific values of q and p.

For intermediate values of k, the welfare implications are more subtle. As in the case of selective sampling, there is a large region of values of q and p for which the researcher is also harmed for some intermediate values of k. However, unlike in the case of selective sampling, the sign of the effect on the evaluator's payoff can be also non-monotonic. In particular, when q is large relative to p, the effect can be negative for relatively small values of k, a phenomenon we have seen occurring only under selective assignment (in that case, for any q and p). This is not too surprising given that, as mentioned above, selective reporting also introduces a dependence between baseline outcomes and treatment status.

We illustrate these conclusions in Figure 5 for the symmetric case $q=p=\frac{1}{2}$, with ΔV_R and ΔV_E denoting the expected utility gains from playing the equilibrium of the game with selective reporting instead of the equilibrium of the non-manipulated experiment. The proposition below provides the statement for general values of q and p.

Proposition 4 *Compared to the non-manipulated experiment, the researcher benefits from selective reporting for*

$$k \in \Big[\mathrm{E}(\beta_{ATE}|Y_t(1)=0), \mathrm{E}(\beta_{ATE}|Y_t^R(1)=1)\Big] \cup \Big[\mathrm{E}(\beta_{ATE}|Y_t(1)=1), \mathrm{E}(\beta_{ATE}|Y_t^R(1)=2)\Big],^{30}$$

[30] The set of values of k where the researcher benefits from manipulation is connected if and only if the first of the two intervals in the union, is empty. This occurs if and only if (34) holds.

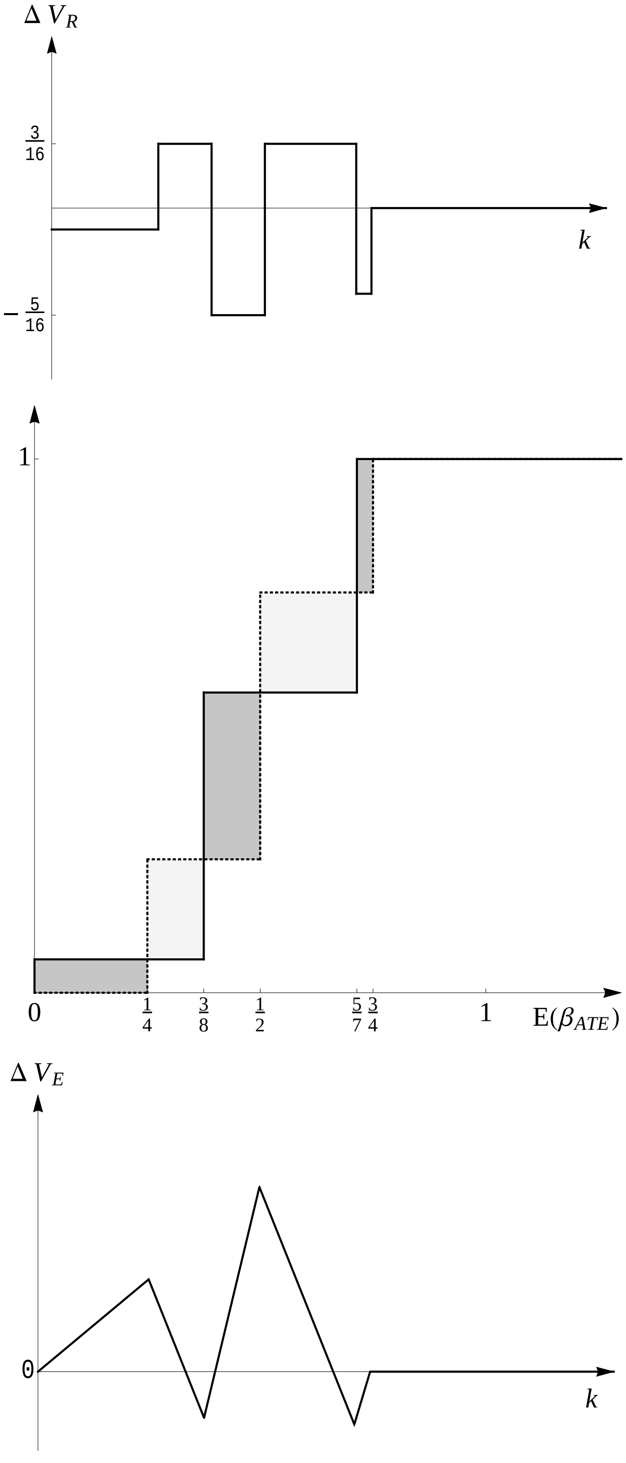

Figure 5: Impact of selective reporting.

and is harmed by it otherwise. Moreover, the evaluator is harmed by selective reporting if

$$k \in \left[\underline{k}^R, \overline{k}^R\right] \cup \left[\hat{k}^R, \mathrm{E}(\beta_{ATE}|Y_t(1)=2)\right],^{31}$$

where

$$\underline{k}^R = \frac{q}{2\left(p+q-pq\right)}, \quad \overline{k}^R = \frac{q\left(1-p\right)\left(1+q-2pq\right)}{2\left(p^2q^2+p+q-3pq\right)}, \quad \textit{and} \quad \hat{k}^R = \frac{1+q\left(1-2p\right)}{2\left(1-pq\right)},$$

and benefits from it otherwise.

7 Conclusion

We conclude by discussing the related literature, summarizing our main messages and suggesting the next steps.

There is little modelling work on persuasion bias in science; see Glaeser (2008) for a discussion of some issues and Lacetera and Zirulia (2011) for a law and economics perspective on scientific misconduct. In an early contribution, Blackwell and Hodges (1957) characterise the experimental design that minimises strategic selection bias in the context of a dynamic reduced-form model positing that selection bias benefits the researcher and harms the evaluator. Once the researcher's information is explicitly modelled, the welfare analysis is much subtler, as our analysis reveals. Berry and Kadane (1997) characterise a situation in which randomised experimental procedure is optimal because of contrasting incentives of different players involved in conducting and interpreting research results.[32]

The models of selective sampling and assignment in Sections 4 and 5 can be seen as equilibrium models of persuasion in which the researcher (sender) has pre-existing private information (hidden information), either about the treatment effect in site $A \in \{M, C\}$ or about the baseline outcome of individual i. Based on that information, the researcher takes an unobservable choice (hidden action), either selecting site M / C or selecting individual i / j. These models have three differences with the models of selective disclosure of Fishman and Hagerty (1990), Glazer and Rubinstein (2004) and Hoffmann, Inderst and Ottaviani (2014).[33] First, our researcher (sender) does not know the realisation of the second random variable (the treatment effect in the second site or the baseline

[31] The set of values of k where the evaluator is harmed is connected if and only if the first of the two intervals is empty, that is, $\underline{k}^R > \overline{k}^R$. This occurs for q small enough relative to p. More precisely, for

$$q < \frac{p\left(1+p^2-p-(1-p)\sqrt{2-2p+p^2}\right)}{4p(1-p)+2p^3-1}.$$

In this case, the qualitative conclusions regarding the impact of manipulation on the evaluator's payoff are identical to those derived under selective sampling. The evaluator is harmed only for large values of k.

[32] See also Tetenov (2014) for a recent development.

[33] See also Dahm, González, and Porteiro (2009) and Henry (2009) for related approaches to disclosure in science.

outcome of the alternative subject), which becomes relevant if the researcher does not select the pre-screened site or assign the pre-screened individual.[34] The second and more important difference is that our evaluator through observation of the experimental outcome obtains some information that the researcher did not have.[35] Finally, in the selective disclosure literature the evaluator's payoff is the sum of the disclosed and the undisclosed component, while the payoff of our evaluator is the average treatment effect.[36]

While in models of optimal persuasion à la Rayo and Segal (2010) and Kamenica and Gentzkow (2011) the sender can freely commit to give any signal to the receiver, our researcher is able to partly affect the signal observed by the receiver only through a hidden sampling/assignment choice.[37] While persuasion benefits the sender in Kamenica and Gentzkow (2011), our researcher faces a commitment problem and thus can be harmed by manipulation. In addition, our baseline for welfare comparison is an ideal RCT rather than the absence of information—thus, our evaluator is an informed receiver.[38]

For the purpose of illustration we developed our arguments in the context of a simple all-binary model. However, we hope to have demonstrated that economics has the potential to help science in a number of ways. First, economic incentives to produce persuasive evidence played a key role in the historical development of experimental methodology. Second, economics offers the tools to understand how persuasion bias affects the interpretation of experimental evidence. Third, economics gives a welfare framework to evaluate the net impact of persuasion bias on the different parties, once an equilibrium situation is reached in which the evaluator properly adjusts the inference for the occurrence of manipulation.

Our equilibrium analysis highlights some subtle welfare implications. Given that more information is present in the system when the researcher strategically uses the information for selection purposes, the evaluator may end up gaining relative to non-manipulation. At the same time, given that the evaluator responds to the researcher's manipulation by raising the standard for acceptance, in equilibrium the researcher sometimes ends up losing from manipulation. In those cases conflicts of interest trap the researcher who would like to commit not to manipulate. In our binary model, for example, with a demanding evaluator and a researcher able to conduct the experiment in a more favourable site, a credibility crisis arises.[39] The researcher loses credibility, and the evaluator ends

[34] Given that the random variables have binary distributions, the researcher would make the same choice as if observing the second variable. This would not be the case if there are more than two realisations, as in Di Tillio, Ottaviani and Sørensen's (2017) model with continuous random variables.

[35] The model of selective reporting in Section 6 can be reinterpreted as a model of selective disclosure where the distribution of the two variables observed by the researcher has three possible values, rather than being binary as in Fishman and Hagerty (1990) or continuous as in Glazer and Rubinstein (2004) and Hoffmann, Inderst and Ottaviani (2014).

[36] A specification of our selective reporting model in which the researcher is perfectly informed about β_M and β_C and reports the highest boils down to Fishman and Hagerty (1990).

[37] See Brocas and Carrillo (2007), Felgenhauer and Schulte (2014) and Henry and Ottaviani (2014) for models of persuasion with sequential acquisition of public information. Here, instead, we focus on a static setting with private information.

[38] See Kolotilin et al. (2015) for a general analysis of optimal persuasion with an informed receiver.

[39] We remark once again that our results are obtained in an all-binary framework, and as such they are meant to be

up with less information.

The natural next step is to generalise the analysis and evaluate the impact of policy measures.[40] In ongoing work we analyse the optimal commitment by the evaluator in the context of a more general informational environment.

A Appendix

We provide the full characterisation of when the researcher benefits from selective sampling, extending Proposition 1 to the case of general values of p and q.

Consider the partition of all possible values of q and p into four regions, which we call cases I, II, III and IV, defined as follows (and represented in Figure 6):

Case I:

$$\frac{1+(1-q)^2}{1+2(1-q)^2} < p < 1.$$

In this case, the conditional expectation of the average treatment effect, given the intermediate outcome under manipulation, is lower than expectation conditional on the lowest outcome under no manipulation: $\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=1\right) \leq \mathrm{E}\left(\beta_{ATE}|Y_t(1)=0\right)$. We illustrate this case in the three panels on the left in Figure 7, for $q=1/2$ and $p=9/10$.

Case II: Either

$$\frac{q^2}{2(1-q+q^2)} < p < \min\left\{\frac{1}{2}, \frac{q^2}{q^2+(1-q)^2}\right\}$$

or

$$\max\left\{\frac{1}{2}, \frac{q^2}{q^2+(1-q)^2}\right\} < p < \frac{1+(1-q)^2}{1+2(1-q)^2}.$$

In this case, we have

$$\mathrm{E}\left(\beta_{ATE}|Y_t(1)=0\right) < \mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=1\right) < \mathrm{E}\left(\beta_{ATE}|Y_t(1)=1\right) < \mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}(1)=2\right).$$

This is the only case featuring a disconnected region of intermediate values of k where the researcher is better off with manipulation—in cases I, III and IV, the region is connected. Case II is illustrated in the three right panels of Figure 7 for $q=1/2$ and $p=8/10$.

illustrative rather than general. Our selective sampling model corresponds to a binary version of Hoffman, Inderst and Ottaviani's (2014) model, with the addition of binary noise.There, manipulation generally benefits the evaluator under logconcavity. The current paper's result that manipulation harms the evaluator at the top is generated by a violation of logconcavity of the left integral of the distribution function at the upper bound of the support. Our selective assignment model is instead closer to Di Tillio, Ottaviani and Sørensen (2017). The fact that manipulation hurts the evaluator at the bottom is similar to a result in that paper, in the case where the support of the distribution is bounded at the bottom.

[40] For example, in economics an active discussion is under way about the pros and cons of pre-analysis plans; see Olken (2015) and Coffman and Niederle (2015). See also Dufwenberg and Martinsson's (2014) proposal of sealed-envelope submissions of experimental results.

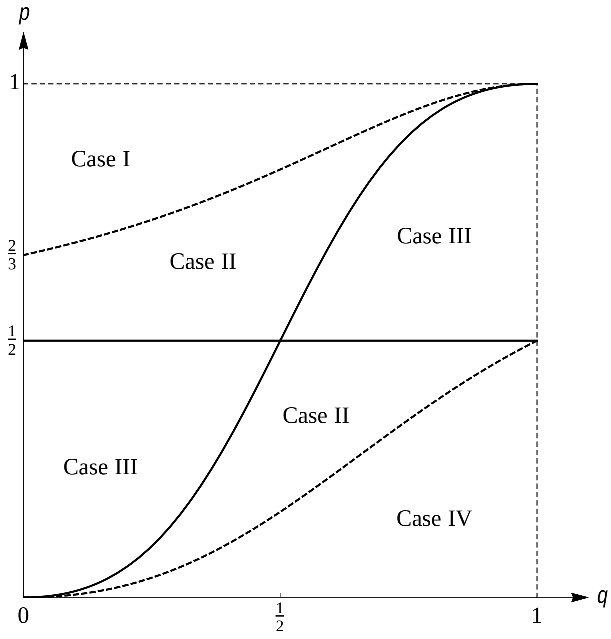

Figure 6: Selective sampling: the four cases.

Case III:

$$\min\left\{\frac{1}{2},\frac{q^2}{q^2+(1-q)^2}\right\}<p<\max\left\{\frac{1}{2},\frac{q^2}{q^2+(1-q)^2}\right\}.$$

In this case the intermediate outcome makes the evaluator more optimistic under manipulation than under no manipulation: $\mathrm{E}\left(\beta_{ATE}|Y_t\left(1\right)=1\right)<\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}\left(1\right)=1\right)$. We illustrate it in the three left panels of Figure 8 for $q=p=7/10$.

Case IV:

$$0<p<\frac{q^2}{2\left(1-q+q^2\right)}.$$

In this case, under manipulation, the conditional expectation of the average treatment effect, given either the intermediate or highest outcome, is lower than the conditional expectation under no manipulation, given the intermediate outcome: we have

$$\mathrm{E}\left(\beta_{ATE}|Y_t\left(1\right)=0\right)<\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}\left(1\right)=1\right)<\mathrm{E}\left(\beta_{ATE}|Y_t^{\beta}\left(1\right)=2\right)<\mathrm{E}\left(\beta_{ATE}|Y_t\left(1\right)=1\right).$$

We illustrate this case in the three right panels of Figure 8 for $q=3/5$ and $p=1/10$.

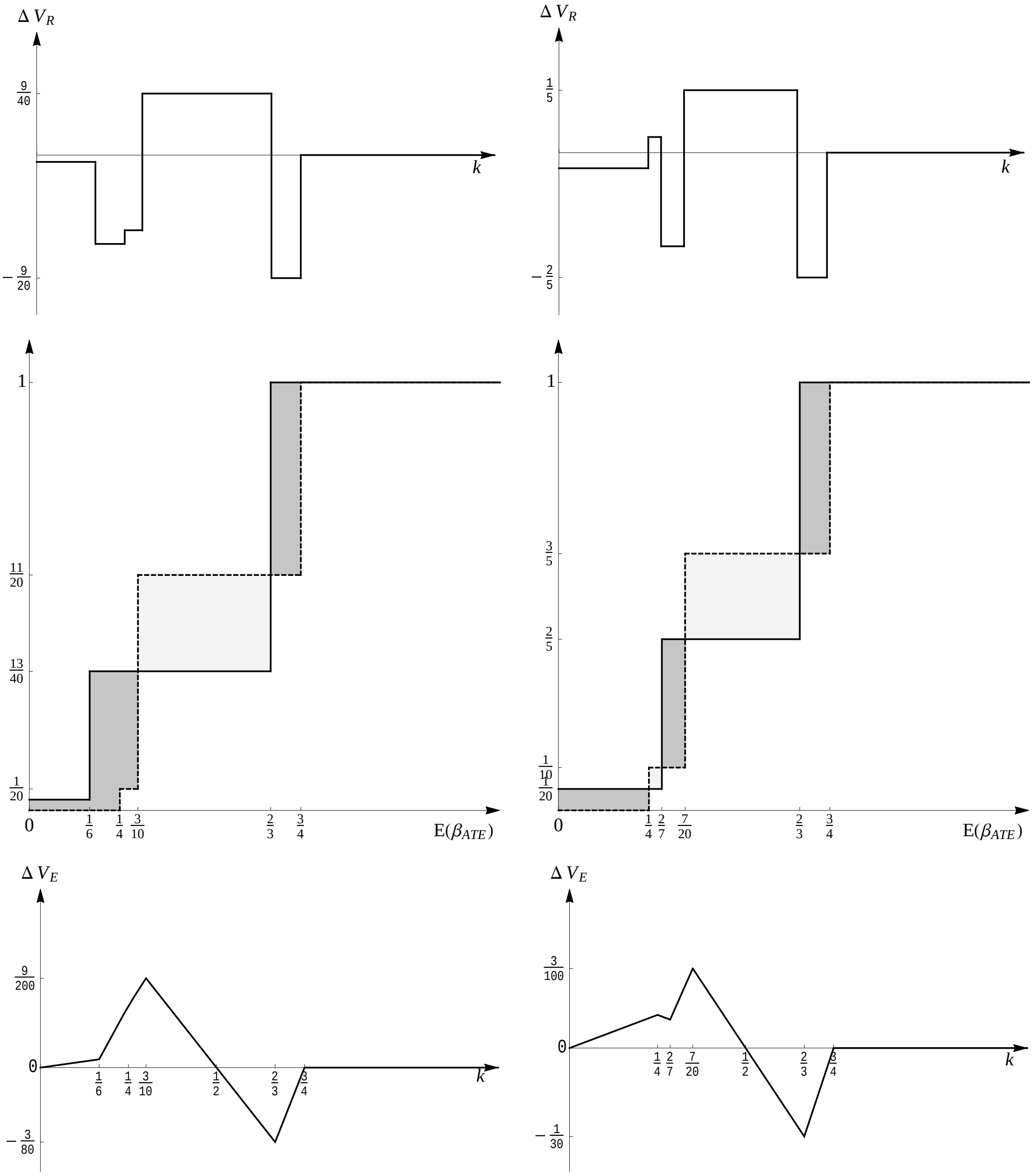

Figure 7: Impact of selective sampling: Cases I (left) and II (right).

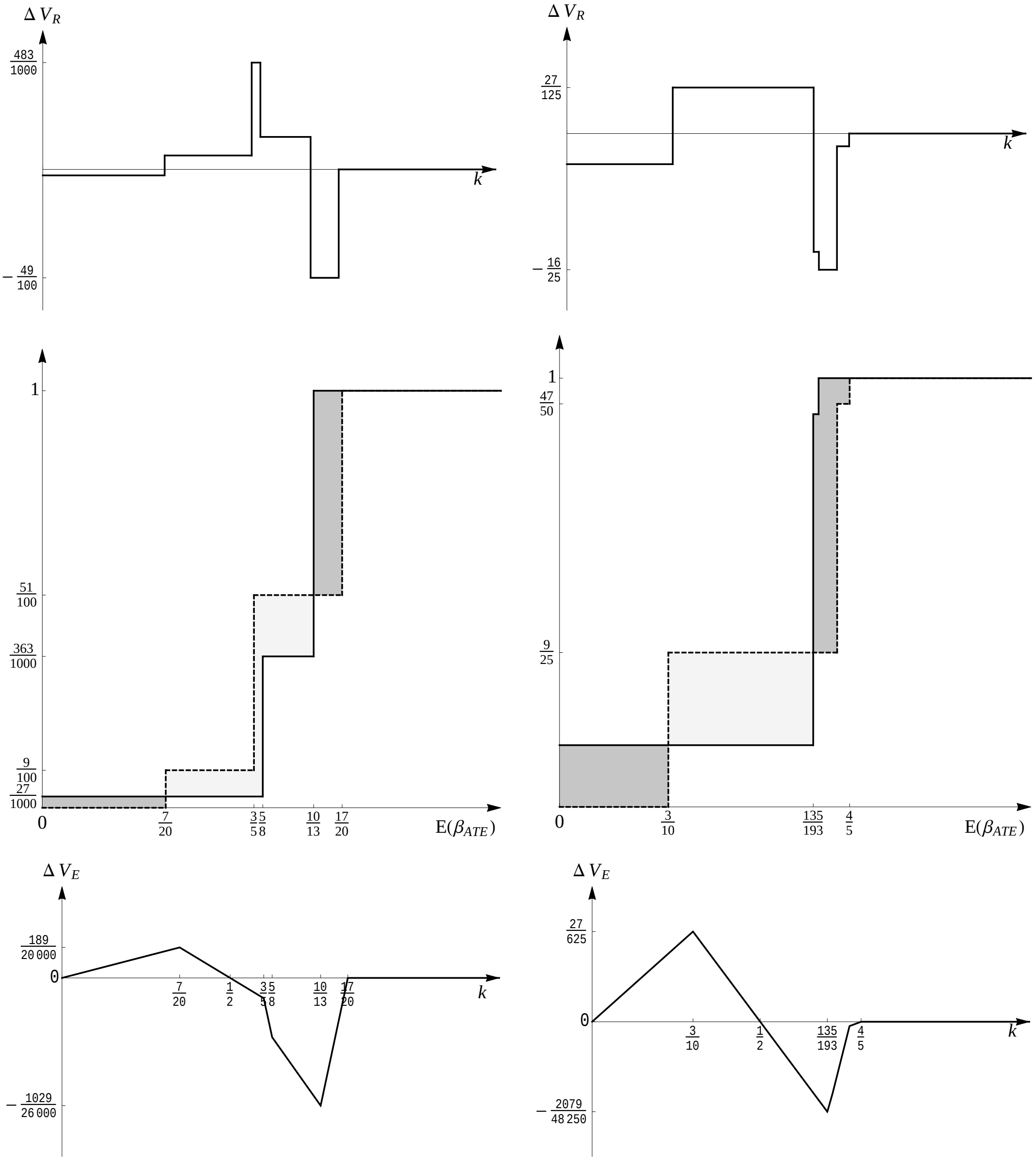

Figure 8: Impact of selective sampling: Cases III (left) and IV (right).

References

Allcott, H. (2015). ‘Site selection bias in program evaluation’, *Quarterly Journal of Economics*, vol. 130, pp. 1117–1165.

Angrist, J.D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton University Press.

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

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

Berry, S.M. and Kadane, J.B. (1997). ‘Optimal Bayesian randomization’, *Journal of the Royal Statistical Society, Series B*, vol. 59(4), pp. 813–819.

Blackwell, D. and Hodges, J.L. (1957). ‘Design for the control of selection bias’, *Annals of Mathematical Statistics*, vol. 28(2), pp. 449–460.

British Medical Journal (1948). ‘The controlled therapeutic trial’, *British Medical Journal*, vol. 2(4582), pp. 791–792.

Brocas, I. and Carrillo, J.D. (2007). ‘Influence through ignorance’, *RAND Journal of Economics*, vol. 38(4), pp. 931–947.

Chalmers, T.C., Celano, P., Sacks, H.S. and Smith, H. (1983). ‘Bias in treatment assignment in controlled clinical trials’, New England Journal of Medicine, vol. 309, pp. 1358–1361.

Chalmers, I. (1999), ‘Why transition from alternation to randomisation in clinical trials was made’, *British Medical Journal*, vol. 319(7221), p. 1372.

Chalmers, I. (2001), ‘Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments’, *International Journal of Epidemiology*, vol. 30(5), pp. 1156–1164.

Chalmers, I., and Clarke, M. (2004). ‘The 1944 patulin trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council’, *International Journal of Epidemiology*, vol. 32, pp. 253–260.

Coffman, L.C. and Niederle, M. (2015). ‘Pre-analysis plans have limited upside’, *Journal of Economic Perspectives*, vol. 29(3), pp. 81–98.

Dahm, M., González, P. and Porteiro, N. (2009). ‘Trials, tricks and transparency: how disclosure rules affect clinical knowledge’, *Journal of Health Economics*, vol. 28(6), pp. 1141–1153.

D'Arcy Hart, P. (1999). 'A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s', *British Medical Journal*, vol. 319(7209), pp. 572–573.

Denton, F.T. (1985). 'Data mining as an industry', *Review of Economics and Statistics*, vol. 67(1), pp. 124–112.

Di Tillio, A., Ottaviani, M. and Sørensen, P.N. (2017). 'Strategic sample selection', working paper.

Dufwenberg, M. and Martinsson, P. (2014). 'Keeping researchers honest: the case for sealed-envelope-submissions', working paper.

Efron, B. (1971). 'Forcing a sequential experiment to be balanced,' *Biometrika*, vol. 58(3), pp. 403–417.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

Fisher, R.A. (1926). 'The arrangement of field experiments', *Journal of Ministry of Agriculture*, vol. 33, pp. 503–513.

Fisher, R.A. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Fisher, R.A. (1936). 'Has Mendel's work been rediscovered?', *Annals of Science*, vol. 1(2), pp. 115–137.

Fishman, M.J. and Hagerty, K.M. (1990). 'The optimal amount of discretion to allow in disclosure', *Quarterly Journal of Economics*, vol. 105(2), pp. 427–444.

Felgenhauer, M. and Schulte, E. (2014). 'Strategic private experimentation', *American Economic Journal: Microeconomics*, vol. 6(4), pp. 74–105.

Glaeser, E.L. (2008). 'Researcher incentives and empirical methods,' in *The Foundations of Positive and Normative Economics: A Hand Book*, eds. A. Caplin and A. Schotter, 300–319. New York: Oxford University Press.

Glazer, J. and Rubinstein, A. (2004). 'On optimal rules of persuasion', *Econometrica*, vol. 72(6), pp. 1715–1736.

Heckman, J.J. (1979). 'Sample selection bias as a specification error', *Econometrica*, vol. 47(1), pp. 153–161.

Head, M. L., Holman, L., Lanfear, R., Kahn, A.T. and Jennions, M.D. (2015). 'The extent and consequences of p-hacking in science', *PLoS Biology*, vol. 13(3), pp. 1–15.

Henry, E. (2009). 'Strategic disclosure of research results: the cost of proving your honesty', *Economic Journal*, vol. 119(539), pp. 1036–1064.

Henry, E. and Ottaviani, M. (2014). 'Reseach and the approval process: the organization of persuasion', working paper, Sciences Po and Bocconi.

Hoffmann, F., Inderst, R. and Ottaviani, M. (2014). ‘Persuasion through selective disclosure: implications for marketing, campaigning, and privacy regulation’, working paper.

Holman, L., Head, M.L., Lanfear, R. and Jennions, M.D. (2015). ‘Evidence of experimental bias in the life sciences: why we need blind data recording’, *PLoS Biology*, vol. 13(7), pp. 1–12.

Ioannidis, J.P.A. (2005). ‘Why most published research findings are false’, *Chance*, vol. 18(4), pp. 40–47.

Ioannidis, J.P.A., Stanley, T.D. and Doucouliagos, H. (2017). ‘The power of bias in economics research’, *Economic Journal*, forthcoming.

Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Princeton University Press.

Jadad, A.R. and Enkin, M.W. (2007). *Randomized Controlled Trials: Questions, Answers, and Musings*, BMJ Books, Blackwell.

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

Kolotilin, A., Li, M., Mylovanov, T. and Zapechelnyuk, A. (2015). ‘Persuasion of a privately informed receiver’, working paper.

Kaptchuk, T.J. (1998). ‘Intentional ignorance: a history of blind assessment and placebo controls in medicine’, *Bulletin of the History of Medicine*, vol. 72, pp. 389–433.

Lacetera, N. and Zirulia, L. (2011). ‘The economics of scientific misconduct’, *Journal of Law, Economics, and Organization*, vol. 27(3), pp. 568–603.

Manski, C. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*, Harvard University Press.

Maniadis, Z., Tufano, F. and List, J.A. (2014). ‘One swallow doesn’t make a summer: new evidence on anchoring effects’, *American Economic Review*, vol. 104(1), pp. 277–290.

Maniadis, Z., Tufano, F. and List, J.A. (2017). ‘To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study’, *Economic Journal*, forthcoming.

Medical Research Council (1944). ‘Clinical trial of patulin in the common cold’, *Lancet*, vol. 2, pp. 373–375.

Medical Research Council (1948). ‘Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation’, *British Medical Journal*, vol. 2(4582), pp. 769–782.

Neyman, J.S. (1923). ‘On the application of probability theory to agricultural experiments. Essay on principles. Section 9.’ Translated in *Statistical Science* (1990), vol. 5(4), pp. 465–480.

Olken, B.A. (2015). ‘Promises and perils of pre-analysis plans’, *Journal of Economic Perspectives*, vol. 29(3), pp. 61–80.

Rayo, L. and Segal, I. (2010). ‘Optimal information disclosure’, *Journal of Political Economy*, vol. 118(5), pp. 949–987.

Rosenberger, W.F. and Lachin, J.M. (2002). *Randomization in Clinical Trials: Theory and Practice*, Wiley.

Rosenthal, R. (1979). ‘The file drawer problem and tolerance for null results’, *Psychological Bulletin*, vol. 86(3), pp. 638–641.

Rothwell, P.M. (2005), ‘External validity of randomised controlled trials: “To whom do the results of this trial apply?”’, *Lancet*, vol. 365, pp. 82–93.

Rubin, D. (1974). ‘Estimating causal effects of treatments in randomized and nonrandomized studies’, *Journal of Educational Psychology*, vol. 66, pp. 688–701.

Rubin, D. (1978). ‘Bayesian inference for causal effects: the role of randomization’, *Annals of Statistics*, vol. 6(1), pp. 33–58.

Rubin, D. (2005). ‘Bayesian inference for causal effects’, Chapter 1 in *Handbook of Statistics, Bayesian Thinking Modeling and Computation*, eds. Dey, D. and Rao, C.R., vol. 25, pp. 1–16. Elsevier.

Schulz, K.F. (1995). ‘Subverting randomization in controlled trials’, *Journal of American Medical Association*, vol. 274, pp. 1456–1458.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995). ‘Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials’, *Journal of the American Medical Association*, vol. 273, pp. 408–412.

Simonsohn, U., Nelson, L.D. and Simmons, J.P. (2014). ‘*P*-curve: A key to the file drawer’, *Journal of Experimental Psychology: General*, vol. 143(2), pp. 534–547.

Tetenov, A. (2015). ‘An economic theory of statistical testing’, working paper.

Torgerson, D.J. and Torgerson, C.J. (2008). *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*, Palgrave MacMillan.

Weisberg, H.I. (2010). *Bias and Causation: Models and Judgement for Valid Comparisons*, Wiley.