# DEBIASING METHODS AND THE ACCEPTABILITY OF EXPERIMENTAL OUTCOMES

David Teira

Dpto. de Lógica, Historia y Filosofía de la ciencia

UNED

Paseo de  senda del rey 7

28040 Madrid

Spain

Phone: (34) 666 379 181

E-mail: dteira@fsof.uned.es

ACKNOWLEDGEMENTS

Research for this paper has been funded by the grant FFI2011-28835. I wrote this paper as a visiting professor of the FOLSATEC graduate program (European Institute of Oncology, Milan). I have received valuable comments from the FOLSATEC students and researchers as well as from Francesco Guala and Jesús Zamora. Two anonymous referees made insightful objections leading to a substantial revision of the paper.

**Abstract**:

Why scientists reach an agreement on new experimental methods when there are conflicts of interest about the evidence they yield? I argue that debiasing methods play a crucial role in this consensus, providing a warrant about the impartiality of the outcome regarding the preferences of different parties involved in the experiment. From a contractarian perspective, I contend that an epistemic pre-requisite for scientists to agree on an experimental method is that this latter is neutral regarding their competing interests. I present two medical experiments (on smallpox inoculation and Mesmerism) in which debiasing procedures such as blinding and data tabulation provided warrants of impartiality that made people agree on the experimental design even if they disagreed on the outcome.

## HOW EXPERIMENTS BEGIN

Most readers probably know well the controversy between philosophers, historians and sociologists on the closure of scientific experiments (e.g., Galison 1987). I am not going to focus here on the closing side of an experiment, but rather on its very beginning: the agreement on its design. Instead of wondering why certain experimental outcomes are accepted, I want to discuss *why certain experimental methods prevail, even when there are controversies on the evidence they yield* . In particular, when it is far from certain that these methods deliver the truth of the matter under study.

More precisely, I want to discuss whether the impartiality of an experimental method plays any role in the acceptance of the setup. I will characterize *experimental impartiality* as follows: an experiment will be impartial if it incorporates

methodological devices preventing the experimenter from manipulating the results according to her interests, in the specific way each device addresses (Teira 2013a, 2013c). Conscious or unconscious, such manipulations generate experimental *biases*: the preferences of the participants in the experiment interfere in the measurement process, altering the result that we would obtain if we replicated the experiment controlling for these preferences. When we detect such interferences, we declare the experimental data *biased*. I will consider an experiment *impartial* to the extent that it incorporates methods that control for the potential sources of bias –e.g., randomization for allocation bias, etc. Therefore, impartiality comes in degrees: the more complete it is our catalogue of potential biases and debiasing methods, the more impartial our experimental procedure will be.

What these methods guarantee is that the experimental procedure will not be manipulated according to the interests of any of the competing parties. Hence, even if the implementation occasionally fails, in the long run these methods make more likely that the truth will emerge from successive replications of the experiment. In the short run, though, we often have no guarantee that an experiment will deliver the truth. Hence, truth by itself may not be enough to ground an agreement on a new experimental setup. Nonetheless, I contend that debiasing methods incorporated into such setup may contribute at this stage to make the design of the experiment acceptable, for a good epistemic reason. I will argue that impartiality is both a pre-requisite for any well-grounded agreement among on methods competing scientists, and also an empirical feature of actual experimental designs that contributes to make them acceptable.

The goal of this paper is to contribute some prima facie evidence for this set of claims about experimental impartiality from a contractarian standpoint (Zamora 2002). First, I am going to present and discuss two "constitutional" episodes in the history of

clinical trials. In both cases, a medical treatment was assessed in novel test where a new debiasing method was implemented for the first time. I want to show how this warrant of experimental impartiality played a practical role in the design and acceptance of these tests.

In testing treatments we find almost by default a constitutive conflict of interests: patients and physicians usually have preferences for or against the treatment; the manufacturer wants it to succeed commercially, its competitors want the opposite. How should a test reconcile all these conflicting perspectives? I contend that the warrant of impartiality provided by our debiasing method contributes to make a practical agreement possible: the stakeholders in a clinical trial do not want their competing interest to influence the outcome, if they are to accept it as fair. From a theoretical standpoint this agreement would be a case in point of a major contractarian intutition: when scientists compete in testing their respective claims (theories, hypotheses, etc), a conceptual pre-requisite for making them agree on an empirical test is that such experiments should be neutral regarding their competing theories. I will defend that the same epistemic property that we characterize in our formal models of scientific agreement, contributes to actual empirical agreement on methods.

The next two sections present two historical treatment tests in which I show, first, that there was an explicit concern for experimental impartiality; and second that the perceived neutrality of the methods implemented in the experiment could have contributed to make it acceptable and widely used. I am going to examine first how the tabulation of data was introduced in 18$^{th}$ century Britain in order to debias the proto-statistical analysis of a trial of inoculation against smallpox. Then I will discuss the introduction of blinding as a screening device in the causal analysis of the therapeutic effects of animal magnetism, in a series of experiments conducted in France before the

Revolution. The reader must bear in mind that my arguments are historically *inconclusive*. I just present a proof of concept: we can interpret these two episodes in the history of medical experimentation from an epistemic stance that does some justice to the actual interests of the agents involved.

In the final section I will sketch a contractarian justification for the epistemic relevance of experimental impartiality. I will also illustrate with some experimental evidence why, as a matter of fact, this sort of impartiality makes more likely that we reach agreements.

## TABULATING INOCULATION DATA

Between the XVII and the XVIII century, a series of outbreaks of smallpox made it the most feared disease in Britain. Unlike other plagues, both rich and poor suffered from the infection and this broad social concern may explain the public controversy over inoculation (Miller 1957, pp. 33-35). There were patients with financial means to buy a cure and there were not many: the best protection against the pox were quarantining the patients or fleeing whenever there was an outbreak (Miller 1957, pp. 35-44). On the other hand, physicians were not always trusted as a profession: a patient often demanded "relief and freedom from pain, little considering about the causes of his illness" (Pender 2006, p. 10), causes about which physicians usually did not agree.

When Lady Montagu introduced in London inoculation, a folk therapy she had come across in Turkey, a public debate over its efficacy started (Miller 1957, pp. 45-133). It seemed indeed a counterintuitive treatment: inoculation put a patient at risk of contracting a disease she may never have had otherwise and die from it. To prove how much she trusted it, in 1721, Lady Montagu had her three-year-old daughter inoculated

and then persuaded the Princess of Wales to inoculate two of her daughters. Many more would follow after her. In the ensuing pamphlet war between supporters and critics of inoculation, the medical profession remained generally incredulous, if not openly critical. They had no theory to explain the purported efficacy of inoculation and they often found it objectionable for both ethical and theological reasons (Miller 1957, pp. 100-110). The empirical part of the controversy hinged on the successful cases of inoculation reported, whether and under which circumstances they constituted a valid proof of the efficacy of the treatment.

We find here one of the first quantitative assessments of a therapy drawing on data gathered to this purpose. The approach was so new that some initially disqualified as "some obscure and improper Calculations scarcely intelligible to any body, or if intelligible, altogether foreign to the Purpose" (Rusnock 2002, p. 68). Mathematics and medicine had then few, if any, intersection points and not many physicians could appreciate the weight of quantitative arguments. However, there was an informal understanding of how an adequate procedure of data recording could correct biased assessments of the success of a therapy. In my view, this understanding motivated the general acceptance of this new experimental method.

Let us focus on the contribution of James Jurin (1684-1750), a physician and Newtonian natural philosopher who served as secretary of the Royal Society between 1721 and 1727[1]. During his tenure, Jurin published a number of pamphlets defending inoculation with proto-statistical arguments (Rusnock 2002, pp. 55-63). He presented his method in the *Letter to the learned Caleb Cotesworth* (Jurin 1723): Jurin argued that a comparison of the figures in the Bills of mortality and the records of inoculated

[1] See Jurin's biography in Jurin 1996, pp. 8-61. John Arbuthnot contributed a similar estimate (Maitland & Arbuthnot 1722).

patients showed that the chances of dying from natural smallpox were bigger than those of dying after inoculation.

Jurin studied the mortality figures for the periods in which smallpox deaths were counted separately (1667-1686 and 1701-1722) calculating that an average proportion of 1 in 14 deaths were due to the disease. Jurin corrected this estimation trying to take into account the high infant mortality caused by the pox and left it in a proportion of 1 in 7 or 8 (Rusnock 2002, pp. 52-53). Jurin prepared another mortality table drawing on the data of several London inoculators during the years 1721 and 1722 and other available reports. He estimated a death ratio for inoculation of 1 of every 91 patients.

Jurin was clear about the goal of his estimates: they were aimed at the British public at large addressing their concerns about smallpox treatments (Jurin 1723, p.7). But why would they be convinced by such figures? As the following citations illustrate, Jurin thought his method to be impartial:

> [I]f the following Extracts and Computations, concerning the comparative Danger of the Inoculated and Natural Small Pox, may be of any Use to your self or to other impartial and disinterested Judges, I shall think my Labour well bestowed. (1723, p. 4)

> And if someone suspects us of Partiality in proceeding after the manner we have done, he need only cast his Eyes upon the second Table (1723, p. 15)

And many of his readers shared this view (see Rusnock 2002, pp. 66-68), among whom I should quote Jurin's friend Thomas Nettleton:

> Your Pieces have been every where well received so far as I can learn, & the more because of the strict Neutrality you observe between the contending Partys. (Jurin 1996, pp. 304-305).

Appearing impartial was certainly necessary in a context where the financial interests of the advocates of inoculation seemed to be affecting their assessment of the treatment. My claim is that Jurin correctly saw the source of this impartiality in a controlled quantitative comparison, where the tabulation of the data prevented any ungrounded optimism in the assessment of the treatment's efficacy. We owe to the remarkable Isaac Massey a very clear description of such biased optimism:

> I remember Mr. Maitland at Child's Coffee-House, when this Practice was just begun at Newgate[2], was as confident and positive of the Success and Security proposed by it, as if he had had Twenty Years Experience without any Miscarriage, which made those who heard him justly suspect, he was more concerned for the Employ than the Successes of it […] (Massey 1723a, pp. 3-4)

According to Massey, Maitland did not have the necessary evidential grounds ("twenty years of experience without miscarriage") to support his confidence in the therapy –this confidence probably originated elsewhere. But how could we estimate which type of evidence (and how much of it) corrected such undue optimism?

Massey produced a subtle methodological pamphlet against inoculation, probably the best piece in this genre (Miller 1957, pp. 100-134). Massey's arguments against inoculation targeted its purported causal mechanism that he considered grounded on a fallacious analogy: the effects of the inoculation are never uniform and depend on the particular circumstances of each patient (Massey 1723a, p. 8). He granted, however, that exposure to the disease may prevent it, if a properly chosen patient

[2] Charles Maitland was a Scottish surgeon who had been Lady Montagu's physician at the British embassy in Constantinople. He inoculated her daughter and became a leading advocate of the treatment (Miller 1957, pp. 71-72).

breathed in a contaminated environment. In order to rule out the efficacy of inoculation, Massey initially called for a controlled observation of its effects:

> [I] make one Request to the Royal College of Physicians; which is, that they would obtain Power from the Government (if they want it) to oblige every Person that shall hereafter be inoculated to have his Name, and Place of Abode, entered in a Register for that Purpose; wherein the Time and Successes, good or bad, should be also Registered, and if afterwards any should live to have the Small Pox, some Care should be taken effectually, that the College be acquainted therewith, and a Memorandum be made of it. (Massey 1723a, p. 21)

He expected to achieve certainty in about 10 years. However, the publication of Jurin's estimates came much earlier, and prompted a quick reaction. Massey published himself a letter questioning the methodology of Jurin's comparison. He invoked an Aristotelian principle sometimes used among physicians[3]: *comparanda non debent habere magnum inter se Differentiam* (1723b, p. 4). Or, as Massey applies it to Jurin's estimates: "To form a just Comparison, and calculate right in this Case, the Circumstances of the Patients, must and ought to be as near as may be on a Par" (1723b, p. 5).

[3] At least, I found through Google Books the following trace: the Dutch physician Johannes Van der Linden (1609-1664) cites it in his *Selecta medica* (Leiden, Johannes Elsevirius, 1656, pp. 440-41) commenting on the fifth book of Hippocrates' *Epidemics*. Van der Linden warned against comparing the resistance to disease of men and women invoking the authority of Aristotle's *Topics* III, 1,116ª5-6: "The question which is the more desirable, or the better, of two or more things, should be examined upon the following lines: only first of all it must be clearly laid down that the inquiry we are making concerns not things that are widely divergent and that exhibit great differences from one another." (Trans. by W. A. Pickard)

Massey saw correctly that a proper comparison of mortality rates requires a fair principle of comparison in order to avoid what we would call today *selection bias*:

> [A]s for Instance, the Weakly, the Rickety, the Consumptive, the Scrophulous, the Asthmatick, and Surfeited Persons, such are all rejected by the Inoculators, as improper for that Operation, and also the Generality of Persons above Majority, and growing into Years (1723b, p. 5)

Hence "the Inoculators" are comparing patients with a good standard of care with the general population, including here many poor patients without access to any treatment; moreover, there is no external check that those inoculated did really suffer from smallpox (Massey 1723b, p. 3). The comparison is biased and Massey ventured his own conjecture if his favourite treatment (breathing infectious air) was fairly compared to inoculation:

> That all Person of equal Ages, Healthiness and Condition of Life under equal Advice, Regimen and Nursing, taken ill of the Small Pox, either with or without Inoculation, the Difference in Success would be but little, yet not to the Advantage of the Inoculated (1723b, p. 14)

In Massey we have indeed a clear understanding of the necessity of a controlled comparison. The controls instantiate in various forms the methodological principle of *comparing like with like*: tabulating data for age, baseline health, treatment, etc. secure proper evidential grounds for the assessment of the efficacy of treatments Such controlled comparisons would correct for the selection bias, whatever its sources (from financial conflicts of interest to patient selection). Tabulating patients taking into account the controlling factors made explicit whether the comparison was fair and warranted the impartiality of the assessment.

Nonetheless Massey seems to have been exceptional in his grasp of the comparative methodology[4]. In my view, Jurin was entitled to claim that his estimates were impartial, since the methodology guiding the construction of his mortality tables had been designed with a view to make the scope of the comparison explicit, rendering transparent to third party analysis any limitations due to subjective biases, whatever their source. And indeed there is evidence to think Jurin was reasoning along proto-statistical principles in which tabulation was explicitly incorporated for bias correction. Andrea Rusnock (2002) points out that Jurin was drawing on the method of numerical tabulation articulated by John Graunt and William Petty. Their *political arithmetic* had been applied only a few decades before to the London Bills of mortality, providing the template for Jurin's calculations. Graunt and Petty had found their own inspiration in Francis Bacon (Rusnock 2002, pp. 16-24), an early advocate of tabulating empirical evidence.

Bacon's tables were not quantitative and, among philosophers of science, they are usually considered as rough inferential tools. Historians of philosophy have pointed out, in turn, how Bacon took them as a proper "laboratory notebooks" aimed at "enhancing the mind's capacity for reflection and minimizing its tendency to distortion" (Muntersbjorn 2003, p. 1138)[5]. Bacon proposed tabulation as a method to control the

[4] According to Andrea Rusnock, "William Douglass, a physician in Boston, Massachusetts, raised similar concerns in his pamphlet *Inoculation of the Small Pox as Practiced in Boston, Consider'd in a Letter to A - S - M.D. & F.R.S. in London* (Boston: J. Franklin, 1722). Douglass argued that the surgeon Zabdiel Boylston had practiced inoculation indiscriminately on the old and young, strong and weak, and that any valid evaluation of the practice must first be made on the young and healthy with a careful record of the outcomes" (personal communication, September 28th 2010). However, Jurin did not reply in print or correspondence to any of them.

[5] See Rossi 1968, pp. 205-207 and Gaukroger 2001, pp. 118-127.

way our senses and memory deal with data: "the mind should not be left to itself, but be constantly controlled (*perpetuo regatur*); and the business done (if I may put it this way) by machines (*per machinas*)"[6]. Recording in writing every symptom observed during regular visits to a patient and then tabulating them would constitute a proper medical record allowing a Baconian physician to relate the nature of the disease, treatments and outcomes, without any subjective distortion (Pender 2006, p. 27).

The impartiality of Jurin's conclusions relied thus on the methodology he adopted for error correction: knowing that our interests could lead us to a selective treatment of evidence, adopting a thorough protocol for the tabulation of data would allow the experimenter –or her audience– to correct her personal equation and secure a comparison of like with like. In my view, Jurin was thus epistemically justified in his claim of impartiality, not because his experiment was actually free from biases, but because he explicitly tried to correct them adopting a method explicitly aimed at it[7].

In sum, Jurin contributed a new method for data analysis in medical experiments, the statistical comparison of epidemiological and experimental data. In order to ground this comparison, he tabulated these data justifying this method for its debiasing properties. In my terms, as a warrant of the impartiality of the analysis. In this regard, I think the controversy around the efficacy of inoculation shows an explicit concern for impartiality as a pre-requisite for finding the truth about treatments. This was my first claim. As to my second claim, we may wonder how persuasive this

[6] F. Bacon. [1620] 2000. *The New Organon*. Edited and translated by Lisa Jardine and Michael Silverthore. New York: Cambridge University Press, p. 28, discussed in Muntersbjorn 2002, p. 1145.

[7] Of course, under a different understanding of impartiality, this claim would not hold. See Teira 2013a, 2013c for a more articulated version of my concept of impartiality

impartiality was, how much it contributed to the adoption of Jurin's approach. About this we only have indirect evidence.

For a start, it may have contributed to the adoption of inoculation in Britain, to the extent that his method helped in making explicit the actual efficacy of the treatment. Initially, it seems as if Jurin's supporters were only in the Royal Society of London; the Royal College of Physicians, for instance, remained silent (Rusnock 2002, p. 44). But in the 1730s physicians had already admitted inoculation as a valid treatment, or at least it was mentioned favourably in nearly every English publication on the topic (Miller 1957, p. 122, pp. 139-46). According to Rusnock (2002, p. 67) and Miller (1957, p. 123), physicians writing 30 years after the publication of the 1723 pamphlet acknowledged his contribution to the acceptance of the therapy. Jurin's estimates tested the toxicity of inoculation and as the number of treated patients increased, the figures became more and more convincing by themselves: 1 death per 500 treated patients by 1765 (Boylston 2002). It was a therapy worth trying, even if there was no follow-up study to see for how long it protected the patient from contagion[8].

Of course, Jurin's audience could not know whether his tabulation method had yielded a reasonably correct estimate just by chance. Apparently, Jurin's claims of impartiality prevailed over Massey's objections and his method was gradually adopted in Britain in the coming decades (Tröhler 2000). There are many possible interpretations of this process but, following Tröhler (2005), I submit that its debiasing properties contributed, as much as any other feature of Jurin's method to its widespread acceptance. My claim, let us recall it, is not that impartiality per se explains the success of Jurin's approach: it is rather that it played a role and, from a contractarian

[8] For a study of the development of inoculation in Britain, see Brunton 1990.

perspective, this could be enough to interpret the adoption of Jurin's method in epistemic terms.

## BLINDFOLDING MAGNETIZED PATIENTS

Let us now address a different kind of bias in medical experiments, arising this time not from the experimenter but from the treated patient. The so-called *placebo* effect is not yet completely understood today, but two centuries ago was systematically appraised for the first time as a source of experimental error in testing therapies: we needed to separate the causal action of the treatment from the effect of the expectations of the patient about such treatment (the placebo)[9]. These expectations may bias a fair comparison between treatments if there is a systematic correlation between expectations and outcomes. Blinding patients as to the treatment they receive breaks this correlation, debiasing the comparison (Teira 2013b)

Blindfolding patients for the administration of treatments was the debiasing method invented by the great chemist Antoine Lavoisier in order to test another publicly controversial therapy seeking official recognition in the prerevolutionary France: *animal magnetism*, also known as Mesmerism. Originally conceived by Franz Anton Mesmer (1734 –1815), a graduate from the Faculty of Medicine in Vienna (1776), it was based on the principle that every disease originated in an imbalance of a universal fluid, that

[9] I identify here the placebo effect with the expectations of the patients, since only these latter would explain why patients improve –and not any active principle in the treatment they receive. This is too gross a simplification, since, on the one hand, most placebo effects can be explained otherwise (e.g., regression to the mean); and, on the other hand, when we observe it, there may be more than expectations at work – see Miller *et al*. (2013) for a discussion. For my case though, the simplification seems acceptable, since the patients expectations were the source of improvement that Lavoisier wanted to control

Mesmer claimed could be controlled and restored to equilibrium by some sort of magnetism. Mesmer arrived in Paris in 1778. Due to the reputation acquired in his first magnetic venture in Vienna, Mesmer's Parisian clinic soon became very popular among the French *élite* (Gillispie 1980). Mesmer probably took himself (or at least, many of his patients and supporters took him) for some sort of scientist and, when his therapy was questioned, Mesmer sought official acknowledgement of animal magnetism by three scientific bodies.

The most successful attempt to get recognition was due to Charles Deslon, a member of the Faculty of Medicine and physician-in-ordinary to the *comte* d'Artois, who became one of Mesmer's first Parisian disciples. Deslon presented animal magnetism to his colleagues, made some of them attend Mesmer's clinic and in 1780 he formally proposed to the Faculty's assembly the conduct of a clinical trial. His sceptical colleagues declined the offer. Mesmer took this badly, threatened to leave France and got an immediate offer on behalf of the king: if a government commission favourably assessed his therapy, he would receive a generous pension and the facilities to create an Institute for Animal Magnetism.

The public success of mesmerism in the pre-revolutionary France can be partly explained if we consider, on the one hand, that Mesmer was challenging the authority of the Church: his animal magnetism would provide a natural explanation of phenomena so far considered extraordinary. On the other hand, the scientific bodies that dismissed his applications appeared to many as illegitimate monopolies on the authority of science, putting it at the service of the Crown (Darnton 1968). This is probably why, in 1784, the Baron de Breteuil, Minister of the King's Household (secretary of state), decided to form a commission in order to examine the claims of Mesmer and eventually

discredit them[10]. Lavoisier made a more general case: if Mesmer was right, everyone could potentially cure using magnetism, subverting thus the institutional organization of medicine in a way a government cannot ignore for its potential consequences for its citizens (Lavoisier 1865, pp. 514-515; Franklin et al. 1784b). Here we find an argument that will reappear time and again in the coming decades, ultimately leading to the creation of State agencies overseeing the pharmaceutical market.

One way or another, the interests in conflict around Mesmerism were so big that the commission needed to appear impartial in order to be credible. But, as a matter of fact, both Lavoisier and Franklin, the most distinguished scientists in the commission, were incredulous even before the actual inquiry started. Franklin had met Mesmer already in 1779 and was never convinced by his therapeutic claims, perhaps because of his own trials with electricity in various types of paralysis (never with lasting effects), perhaps because he was well aware of how we misinterpret spontaneous remission: "there being so many Disorders which cure themselves and such a Disposition in Mankind to deceive themselves and one another on these Occasions" (Smyth 1905, p. 183). Lavoisier shared this view: since it is difficult to tell apart the influence of the treatment from all other factors, we need to accumulate data on the effects of the former in order to make a probabilistic assessment. (Lavoisier 1865, p. 599)

[10] Actually there were two commissions, since another one was arranged at the Society of Medicine. We will just pay attention to the first one, chaired by Franklin, where the blinding procedure was originally employed. We will draw on these original sources: their report (Franklin et al. 1784a), the summary of the report presented at the Academy of Sciences (Franklin et al. 1784b), the various manuscripts prepared by Lavoisier on the topic collected in his *Memoir sur le magnétisme animal* (Lavoisier 1865) and Deslon's (1784) reply to the report.

After reading summaries of Mesmer's doctrine, Lavoisier conjectured, as some others had done before him (Lavoisier 1865, p. 505), that what he called the patient's "imagination" might explain the immediate effects of the therapy and planned several experiments in order to test this possibility.

The commission, in which there were three physicians together with Lavoisier and Franklin, was formally summoned in April 1784 by the Baron of Breteuil and spent the next four months conducting experiments in Paris with the assistance of the aforementioned Charles Deslon, who had quarrelled with Mesmer and established an independent clinic in 1781. Right in the acceptance letter to Breteuil, the commission announced that they would not judge the therapeutic effects of animal magnetism, since they could not expect to gather enough data to reach a firm conclusion (Lavoisier 1865, p. 500).

The commission was going to assess instead whether animal magnetism had any physical reality, under the principle that it "can exist without being useful, but it cannot be useful if it does not exist" (Franklin et al. 1784b, p. 8). However, they acknowledged that the only empirical effects discernible were those manifested in patients undergoing magnetic therapy. Namely, convulsions and minor sensations (such as warmth or coldness) following minimal intervention from the magnetizer (light touches, signs, gestures). The experiments focused on these effects, trying to isolate their purported causes. Here is where the innovative design came to set a standard for further research.

Lavoisier's assumption in designing the inquiry was that the causes could be either physical (like animal magnetism) or "moral" (1865, p. 510). He knew how to proceed with the former, and just a series of preliminary observations in about 14

patients convinced the commission that they had no discernible effect[11]. Yet, there was no previous scientific research about the physical effects of moral causes (Franklin et al. 1784b, p. 11). Nonetheless, Lavoisier drew on his laboratory experience in chemistry, decomposing and recomposing substances (Franklin et al. 1784b, p. 10): the test of animal magnetism would "analyze" the physical and moral causes, pretending that both are acting, when only one of them actually does. The "synthesis" of the convulsions in each case would prove the success of the analysis (Lavoisier 1865, p. 510). In order to make this experimental setup work uncontaminated by the preferences of the patients, Lavoisier came up with a brilliant debiasing method: blindfolding them. Predictably, when the experiment was actually carried out, there were blinded patients who suffered convulsive crisis when they were told they were being magnetized (and they were not) and remained calm when they were in fact magnetized but nobody told them[12].

The commission tried to explain the psychological and physiological mechanisms at play in the convulsions (Franklin et al. 1784a, pp. 48-63). As to the former, they claimed that imagination alone or combined with the effects of touching and imitation accounted for the immediate effects of the therapy, but they just illustrated these mechanisms with various examples, without formal definitions. The physiological mechanisms explaining the convulsions were discussed in order to assess their potential effects on the patient's health: the Commission warned these could be negative and advised to proscribe magnetic therapies.

[11] The way the commission selected the patients and dismissed their testimonies was the major source of objections against the report, since it often presupposed the hypothesis that Lavoisier sought to test: the patients were under the influence of the experimenter's expectations: see Lynn and Lilienfeld 2002.

[12] For the actual report of the experiments, see Franklin et al. 1784a, pp 31-48. The plan is laid out in Lavoisier 1865, pp. 511-513.

Of course, the commission was later criticized on both accounts, since their conjectures were unsupported by the available evidence (in their experiments or elsewhere). But leaving these objections aside, what earned the Commission a place in the History of clinical trials was their finding that there were psychological mechanisms at play by which a doctor could act on a patient, causing physical effects that distorted the tests of a purportedly physical therapy (Franklin et al. 1784b, p. 15). Even if the precise articulation of these mechanisms was unknown, the Commission put forward a comparative design in which the action of these mechanisms was separated, securing it with blindfolding as a debiasing method.

Unlike with Jurin's tabulation method we do not find an explicit concern for impartiality in the texts of Lavoisier. However, if we recall my definition of experimental impartiality, this is what his method brought about: blindfolding was indeed a methodological device preventing the experimenter (and the patient) from manipulating the results according to her interests. Lavoisier was perfectly aware of how the imagination of the patients could be manipulated to create the illusion of a cure: blindfolding patients prevented the mesmerist from interacting directly with them, breaking off any systematic connection between the interests of the therapist and the expectations of patients.

Unlike with Jurin, we find here a clear acknowledgement of the epistemic virtues of blindfolding independently of the quality of the outcome of the experiment. The conclusions of the Commission prompted controversy, since the interests in favour of Mesmerism were big, and its supporters immediately contested it. In my view, it is remarkable that the debiasing properties of blindfolding were rarely contested in the pamphlet war that followed the publication of the Commission's report –20.000 copies were sold of this latter alone.

Magnetic therapy was cultivated in France for several more decades, prompting the conduct of new experiments in order to grasp its efficacy[13]. Very few of these latter questioned the necessity of blindfolding the patients, since their expectations regarding the therapy could bias the results of the experiment: Deslon himself seems to have admitted it, according to the report –he tried later to qualify his admission though (Franklin et al. 1784, p. 60). What the Mesmerists questioned was that the patients' "imagination" alone could explain the healing of so many patients treated with animal magnetism. In the pamphlets the possibility of a spontaneous remission was rarely considered, partly because the recovery rate of magnetotherapy did not seem inferior to many other legitimate medical treatments. Hence, the Mesmerists claimed the Commission should have conducted experiments with patients who were naturally blinded and had yet recovered[14].

The Commission was accused, of course, of professional bias by the way they designed the experiment[15]. In any case, the principle that trials of animal magnetism should be blinded was rarely questioned. As Ted Kaptchuk (1998, p. 397) puts it:

> Debunkers and advocates alike quickly adopted the new blind assessment method to prove their points of view, and it became intrinsic to the entire controversy surrounding the nineteenth-century medical and extramedical mesmeric movement. In cloistered academic laboratories and on stages before hundreds, magnetic healers and itinerant entertainers were challenged to cure, detect, or perform wondrous feats with

[13] For a survey of ensuing research on Mesmerism, see Bertrand 1826 and Burdin and Du Bois 1841. The main arguments of the pamphlet war are summarized in Pattie 1994.

[14] The clearest presentation of these arguments I found is Servan 1784. See pp. 109-111 for an alternative plan for the inquiry.

[15]See, e.g. M. G. C*** 1784, pp. 13-14.

practitioners and/or subjects blindfolded. A cottage industry of blind assessment developed.

I suggest interpreting this general acceptance of blinding as an acknowledgment of the necessity of an impartial trial. After all, the supporters of Mesmerism might have contested this debiasing method, arguing –as many others did later– that the perceived effects of the therapy were part of the treatment. Lavoisier's approach prevented the preferences of patients and therapists from having an influence on the trial outcome, making the experiment more credible than it otherwise had been.

As a reviewer observes though, it is true that, despite being a relatively old technique, blinding has been very unevenly used by experimenters across disciplines[16]. This would imply that the concern for experimental impartiality is not as wide as we would like to think. This may be true: my claim so far is that impartiality becomes a prominent concern only when we have big conflicts of interest at stake in a trial that prevent parties from reaching an easy agreement on its outcome. Paradigmatically this is the case of medical experiments, where, for instance, patients are often willing (or not) to use treatments disregarding significant pieces of evidence about their actual effects. Of course, impartiality is neither a necessary nor a sufficient condition for an actual agreement on an experimental outcome. What I claim is that, from an epistemic standpoint, such a consensus is more justified when there are underlying warrants of impartiality such as the debiasing procedures we have been discussing. We will try to motivate this further in the conclusion.

[16] As to the former, blinding was already used for the test of exorcisms in the 16th century: see Kaptchuk et al. 2009. For the sometimes poor understanding of the virtues of blinding among experimenters, see, for instance, the discussion of the predesignation rules among experimental physicists in Staley 2002

## IS THERE A GENERAL CONCERN FOR EXPERIMENTAL IMPARTIALITY?

The two case studies present *prima facie* evidence for two claims. On the one hand, in controversial experiments, there is an explicit concern for debiasing methods as warrants of the impartiality of the outcome. On the other hand, these methods contribute to close the controversy on the hypotheses tested. From a philosophical standpoint, there is a simple interpretation of both cases: in comparative experiments, such as clinical trials in medicine, a bias is just a confounding factor breaking the causal parity of both groups (regarding every other factor but the treatments administered). This is a perfectly valid interpretation, since this is an actual contribution of any debiasing method. I want to supplement it with a different approach, focusing on why these methods were widely adopted.

From most philosophers of science, a simple concern for truth is enough to explain it, since experimenters would be just truth seekers and debiasing methods contribute to this search. But historians and sociologists have shown that in experiments such as those discussed in the previous sections, there is no general agreement on what counts as a confounding factor, since the mechanisms underlying each treatment were not clear themselves –see, e.g., Collins and Pinch (2005), pp. 84-110. The truth about each therapy only emerged in the long run, after the evidence accumulated. But this accumulation was only possible because there was, at least, an agreement on the methods for gathering evidence. So how and why did scientists reach this consensus?

I contend that, from an epistemic standpoint, *this agreement can be interpreted as grounded on the impartiality of the experiment, warranted by the introduction of debiasing methods such as tabulation or blinding*. After all, scientists may disagree on the unknown confounding factors in an experiment, but they usually acknowledge that

their own interests are a known confounder that should be controlled for. We can vindicate thus the Baconian intuitions behind Jurin's tabulation method. Debiasing methods guaranteed that the outcome of the test was not contaminated by the preferences of the experimenters (or any other participant in the experiment). This is what Lavoisier achieved blindfolding the patients. Despite the controversy on the effectiveness of the therapies, both debiasing procedures (tabulation and blinding) were widely adopted by the contending parties. And, in my view, they were epistemically justified in such consensus on experimental methods.

The epistemic relevance of controlling for these preferences can be defended from a contractarian perspective (Zamora Bonilla 2002), as I have tried to argue elsewhere (Teira 2013a). The core intuitition of this approach is as follows: Let us imagine a community of self-interested scientists, partly motivated by finding truths, and partly by more mundane interests such as, for instance, the success of their careers. They are in competition with each other for achieving this success, so when they test their respective claims (theories, hypotheses, etc.), they have every incentive to contest each other's outcomes. In a winner-takes-it-all scenario, the success of one of these scientists may imply the failure of the rest of them, so they can prevent it by rejecting as flawed everybody else´s tests. However, if none of them accepts each other's outcomes, there will be no winner in the race for success: A researcher seeking to increase her professional accomplishments can only succeed if her peers accept her results.

From a contractarian standpoint, even if every member of this community of self-interested scientists has an incentive to promote the experimental methods most favorable to her own theory, they need to agree on a set of shared testing standards, *so that at least one of the competing parties can win*. A pre-requisite for the shared acceptance of any of these testing standards is that they are impartial in the sense

discussed above: they should incorporate methodological devices preventing the experimenter from manipulating the results according to her interests, not giving anyone a better chance of finding the result they are after.

This intuition can be substantiated in a formal economic model of scientific agreement (e.g., Ferreira and Zamora Bonilla 2006), showing under which range of circumstances this latter is possible. If we accept a characterization of scientists as self-interested agents maximizing a utility function, striking a trade-off between truth and social scuccess, experimental impartiality will be a pre-requisite for their agreement on a testing standard, and the subsequent experimental results. In the two cases studies presented above we saw how the concern for debiasing methods was rooted in a combination of epistemic and practical interests. Physicians cared about the true effect of a treatment, but they were also concerned by the money they could make or lose depending on the outcome of a trial. Debiasing methods provided a warrant of impartiality that, I submit, would contribute to close any controversy on the test of a treatment: if we accept *ex ante* the fairness of statistical tabulations or blinding devices, we should not contest as biased, *ex post*, the experimental outcome. We have developed this argument in a systematic fashion elsewhere –see Teira 2013a,b.

However, from an empirical standpoint, and in order to persuade historians and sociologists, we need some evidence that the agents involved in actual experiments have this taste for impartiality presupposed in our contractarian epistemology. We want to close this paper showing that there is as well *prima facie* evidence for such a taste, at least if we accept an analogy between experiments and fair decision procedures. As we have seen in the previous two sections, medical experiments are generally perceived by all the concerned parties as a *decision procedure* on the efficacy of a therapy. This efficacy is usually disputed and, ideally, the experiment should end the controversy and

make us agree. However, the agreement involves different costs and benefits for the concerned parties: for instance, the producer of the therapy (or his competitors) may earn or lose the money patients will be willing to pay for it. As the two cases presented illustrate, the concerned parties tend to accuse each of other of partiality: their personal interests in the outcome of the experiment may bias it as a *decision procedure*, making it unfair.

When it comes to distributing costs and benefits in any context, the perceived fairness of our decision procedure seems to contribute to the acceptance of the outcome when it is adverse to our interests. To the extent that scientists compete, there are winners and losers in every experiment and the implementation of debiasing procedures will increase the perceived fairness of the test, making its outcome more palatable to those who lose. Psychologists have been studying throughout the last four decades the individual reactions to the fairness of a third party decision. The general approach implemented in many different experimental settings, from the field to the laboratory, assumed that each decision follows some sort of procedure and generates an outcome for the participant whose reaction is studied. The experiments usually aim at disentangling how this reaction depends on the perceived justice of the procedure, on the one hand, and the perceived justice of the outcome, on the other hand.

The following generalizations seem to hold (I quote from Brockner and Wisenfeld 1996, p. 191):

- When outcomes are unfair or have a negative valence (e.g., losses), procedural justice is more likely to have a direct effect on individuals' reactions
- When procedural justice is relatively low, outcome favourability is more apt to be positively correlated with individuals' reactions

- The combination of low procedural fairness and low outcome favourability engenders particularly negative reactions

I conjecture that debiasing procedures operate in this same manner inside and outside scientific experiments, and this is why they contribute to make the outcome of controversial tests more acceptable. Think, for instance, of randomization: in comparative experiments, a randomized allocation of treatments prevents the experimenter from distributing them according to her own preferences. It is a warrant of impartiality. But a randomized allocation is just a fair lottery, where, for instance, every patient in a trial has the same probability of receiving any of the treatments under analysis. We have experimental evidence that randomization, as a decision procedure, increases the perceived fairness of any distribution process.

Consider now this economic experiment conducted by Bolton and co-authors (2005). It is an Ultimatum game in which one of the participants (the proposer) must choose how to share a given amount of money and the other participant (the responder) must accept (a) or reject (r) the offer. For instance, the proposer splits 2000 units, keeping 1800 for himself and 200 for the responder: if the latter accepts (a) this is what they have; if he rejects it (r), they both receive 0. Bolton and his co-authors arranged a randomized ultimatum in which a random draw decided which of the three splits was offered to the responder, with three different probability assignments (figure 1)

**Figure 1: Ultimatum game (Bolton et al. 2005)**

In ASYM the dice are loaded for the most favourable payoff to the proposer (proposal C). In SYM98, the loaded proposal is the equal (fair) payoff. In SYM34 the three payoffs are almost equiprobable. Bolton et al. (2005, p. 1065) observed that the fair lottery (SYM34) prompted the responder to accept an unfair offer (1800-200) just as often as in SYM98, where the fair split (1000-1000) is probabilistically loaded

(0,98). The unfair offer (1800-200) was rejected significantly more often in ASYM than in SYM34: when the dice are loaded for an unfair distribution, we don't find it as acceptable as when the lottery is fair.

If the analogy between experiments and distribution processes holds, randomization not only guarantees the impartiality of an experiment, making the outcome neutral regarding the interests at stake; it makes the costs involved in this outcome more acceptable for those who lose (e.g., scientists who see their hypotheses fail). If there is indeed an empirical preference for impartial decision procedures, this is a lever to apply the contractarian approach to actual episodes of controversy about experimental design, such as those discussed in the previous sections. Philosophers of science might see this as just another instance of old epistemic rationality prevailing in real life. Historians and sociologists may be persuaded instead because, in our contractarian approach, we do not presuppose much rationality: scientists seeking their own interest with a moderate taste for fairness in the distribution of actual costs may reach methodological agreements such as those we observe in real life.

## CONCLUDING REMARKS

We have seen two historical illustrations of experiments about medical treatments in which, despite serious conflicts of interests between the concerned parties, they reached an agreement about the debiasing methods implemented in the tests. Tabulating data about smallpox inoculation allowed the experimenter to monitor the factors influencing the comparative efficacy of the treatment and allowed his audience to check whether such comparisons were fair. Blindfolding patients in the Mesmerism trials allowed the experimenter to control for their expectations and tear them apart from the actual treatment effect.

In both cases there was no clear understanding of the biases that could interfere with the test, but the debiasing methods implemented seemed to convinced the concerned parties that the experiment had been impartial enough. These methods provided a warrant about the non-manipulability of the evidence according to anyone's preferences. If the experimenter commits himself to tabulate his experimental data or have his subjects blindfolded, he loses freedom to fiddle with the test in his own interest. From a contractarian perspective, we can vindicate this understanding of impartiality as non-manipulability as an epistemic pre-requirement for reaching a scientific consensus on any experimental method. There is evidence to think that this preference for debiasing procedures as warrants of fairness actually plays a role in our actual decision-making. Hence, from a contractarian approach we can make sense of the epistemic impact of debiasing methods in the history of clinical trials.

## REFERENCES


Bertrand, Alexandre-Jacques-François. 1826. *Du magnétisme animal en France, et des jugements qu'en ont portes les sociétes savantes ... suivi de considérations sur l'apparition de l'extase, dans les traitements magnétiques*. Paris: Bailliére.

Bolton, Gary E, Jordi Brandts, and Axel Ockenfels. 2005. "Fair Procedures: Evidence from Games Involving Lotteries." *The Economic Journal* 115 (506):1054-1076.

Brockner, J., and W. N. Wiesenfeld. 1996. "An Integrative Framework for Explaining Reactions to Decisions: Interactive Effects of Outcomes and Procedures." *Psychological Bulletin* 120 (2):189-208.

Brunton, Deborah. 1990. *Pox Britannica: smallpox inoculation in Britain, 1721-1830*.

Burdin, Charles, and E. Frederic Dubois. 1841. *Histoire académique du magnétisme animal, accompagnée de notes et de remarques critiques. Sur toutes les*

*observations et expériences faites jusqu'a ce jour*. Paris: Chez J.-B. Bailliére Libraire de l'Académie royale de médecine ...;

Chalmers, Iain, and Robert Matthews. 2006. "What are the implications of optimism bias in clinical research?" *Lancet* 367 (9509):449-50.

Collins, H. M., & Pinch, T. J. 2005. *Dr. Golem: How to Think About Medicine*. Chicago: University of Chicago Press.

Darnton, Robert. 1968. *Mesmerism and the end of the Enlightenment in France*. Cambridge, Mass.: Harvard University Press.

Deslon, Ch. 1784. *Observations sur les deux rapports de MM. les Commissaires nommés par Sa Majesté, pour l'examen du magnétisme animal*. Paris: Chez Clousier.

Ferreira, J. L., and Jesús Zamora. 2006. "An Economic Model of Scientific Rules." *Economics and Philosophy* 22:191–212.

Franklin, B., and et alii. 1784a. *Rapport des Commissaires chargée par le Roi, de l'examen du magnétisme animal*. Paris: L'Imprimerie Royale.

Franklin, B., and et alii. 1784b. *Exposé des expériences qui ont été faites pour l'examen du magnétisme animal / Lû à l'Académie des Sciences, par M. Bailly, en son nom et au nom de Mrs. Franklin, le Roy, de Bory et Lavoisier, le 4 septembre 1784.* Paris: Imprimerie Royal.

Galison, Peter. 1987. *How experiments end*. London-Chicago: The University of Chicago Press.

Gaukroger, Stephen. 2001. *Francis Bacon and the transformation of early-modern philosophy*. Cambridge: Cambridge University Press.

Gillispie, Charles Coulston. 1980. *Science and polity in France at the end of the old regime*. Princeton, N.J.: Princeton University Press.

Jurin, James, and Caleb Cotesworth. 1723. *A letter to the learned Caleb Cotesworth ... Containing, a comparison between the mortality of the natural small pox, and that given by inoculation ... To which is subjoined, an account of the success of inoculation in New England; as likewise an extract from several letters concerning a like method of communicating the small pox, that has been used time out of mind in South Wales*. London: W. and J. Innys.

Jurin, James. 1996. *The correspondence of James Jurin (1684-1750): physician and secretary to the Royal Society*. Edited by Andrea Alice Rusnock. Amsterdam; Atlanta, Ga.: Rodopi.

Kaptchuk, Ted J. 1998. "Intentional Ignorance: A History of Blind Assessment and Placebo Controls in Medicine." *Bulletin of the History of Medicine* 72 (3):389-433.

Kaptchuk, T. J., Kerr, C. E., Zanger, A. 2009. "Placebo controls, exorcisms, and the devil". *The Lancet*, 374 (9697): 1234 – 1235.

Lavoisier, A. L. 1865. "Sur le magnétisme animal." In *Oeuvres de Lavoisier, vol. III*, 499-527. Paris: Imprimerie Impériale.

Lynn, Steven Jay, and Scott Lilienfeld. 2002. "A critique of the Franklin commission report: Hypnosis, Belief and suggestion." *International Journal of Clinical and Experimental Hypnosis* 50 (4):369 - 386.

M. G. C***, 1784. *Observations sur le Rapport des commissaires chargés par le Roi de l'examen du magnétisme animal*. Vienne: s.n.

Maitland, Charles, and John Arbuthnot. 1722. *Mr. Maitland's account of inoculating the small pox vindicated [by himself], from Dr. Wagstaffe's misrepresentations of that practice, with some remarks on Mr. Massey's sermon*. London: J. Peele.

Massey, Isaac. 1723a. *A short and plain account of inoculation. With some remarks on the main arguments made use of to recommend that practice, by Mr. Maitland and others*. The second ed. To which is added / ed. London: W. Meadows.

Massey, Edmund. 1723b. *Mr. Maitland's account of inoculating the small pox ...: To which is added, a postscript confirming the success of this practice, from Mr. Massey the apothecary's pamphlet on the subject. And a word to the Reverend Mr. Massey on his vindication of his sermon*. The second edition. ed. London: Printed and sold by J. Peele.

Miller, Genevieve. 1957. *The adoption of inoculation for smallpox in England and France*. Philadelphia,: University of Pennsylvania Press.

Miller, F. G, Colloca, L., Crouch, R.A., Kaptchuk, T. J. 2013. *The Placebo: A Reader*. Baltimore: Johns Hopkins University Press.

Muntersbjorn, Madeline M. 2003. "Francis Bacon's Philosophy of Science: *Machina intellectus* and *Forma indita*." *Philosophy of Science* 70 (5):1137-1148.

Pattie, Frank A. 1994. *Mesmer and animal magnetism: a chapter in the history of medicine*. Hamilton, N.Y.: Edmonston Pub.

Pender, Stephen. 2006. "Examples and experience: on the uncertainty of medicine." *The British Journal for the History of Science* 39 (01):1-28.

Rossi, Paolo. 1968. *Francis Bacon: from magic to science*. London,: Routledge & K. Paul.

Rusnock, Andrea Alice. 2002. *Vital accounts: quantifying health and population in eighteenth-century England and France*. Cambridge: Cambridge University Press.

Servan, J.-M.-A. 1784. *Doutes d'un provincial, proposés à Messieurs les médecins-commissaires, chargés par le Roi, de l'examen du magnétisme animal*. Lyon-Paris: Chez Prault.

Smyth, Albert Henry, ed. 1905. *The writings of Benjamin Franklin.* Vol. 9. New York: Macmillan.

Staley, Kent W. 2002. "What Experiment Did We Just Do? Counterfactual Error Statistics and Uncertainties About the Reference Class". *Philosophy of Science* 69: 279-299.

Teira, David. 2013a. "A contractarian solution to the experimenter's regress", *Philosophy of science* 80: 709-720.

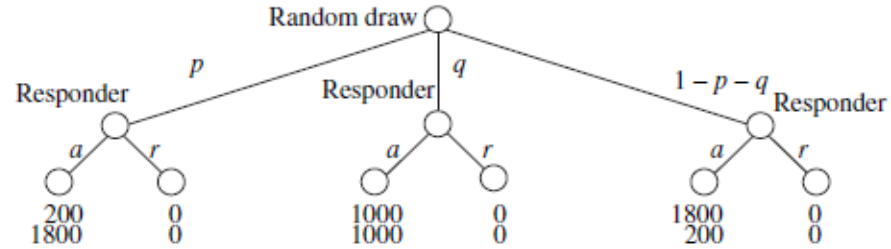
Teira, David. 2013b. "Blinding and the non-interference assumption in field experiments", *Philosophy of the Social Sciences*, 43.3: 358-372

Teira, David. 2013c . "On the impartiality of British trials", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 44.3: 412-418.

Tröhler, Ulrich. 2000. *"To improve the evidence of medicine": the 18th century British origins of a critical approach*. Edinburgh: Royal College of Physicians of Edinburgh.

Tröhler, Ulrich. 2005. "Quantifying experience and beating biases: A new culture in eighteenth-century British clinical medicine." In *Body Counts: Medical Quantification in historical and sociological perspective*, edited by G. Jorland, A. Opinel and G. Weisz, 19-50. Montreal, London, Ithaca: McGill University Press.

Zamora, Jesús. 2002. "Scientific Inference and the Pursuit of Fame: A Contractarian Approach." *Philosophy of Science* 69:300-323.




Treatments: (1) ASYM: $(p, q, 1-p-q) = (0.01, 0.01, 0.98)$, (2) SYM 98: (0.01, 0.98, 0.01), and (3) SYM34: (0.33, 0.34, 0.33)

Figure 1