Why Experimental Balance is Still a Reason to Randomize

Marco Martinez & David Teira

Abstract

Experimental balance is usually understood as the control for the value of the conditions, other than the one under study, which are liable to affect the result of a test. We will discuss three different approaches to balance. 'Millean balance' requires to identify and equalize *ex ante* the value of these conditions in order to conduct solid causal inferences. 'Fisherian balance' measures *ex post* the influence of uncontrolled conditions through the analysis of variance. In 'efficiency balance' the value of the antecedent conditions is decided *ex ante* according to the efficiency they yield in the estimation of the treatment outcome. Against some old arguments by John Worrall, we will show that in both Fisherian and efficiency balance there are good reasons to randomize the allocation of treatments, in particular when there is no agreement among experimenters as to the antecedent conditions to be controlled for.

1 Why There Was No Cause to Randomize

It is almost 20 years since John Worrall published 'What evidence in evidence-based medicine?' (Worrall, 2002), the first in a series of papers contesting the purported epistemic superiority of randomized clinical trials (RCTs) (Worrall [2007a], [2007b], [2008]). Hierarchies of evidence in medicine usually placed RCTs near the top of the pyramid, under the widely shared assumption that randomized tests provided the best

evidence about the safety and efficacy of medical treatments, since they controlled better for biases than any other approach. Worrall challenged this assumption, with a battery of arguments arguing that randomization did not control better for balance than carefully controlled non-randomized experiments.

Worrall's assessment of randomization is now a mainstream view among philosophers of medicine -see, for instance, (Solomon, Simon, & Kincaid [2017]). Yet, in the last two decades, trust in randomized experiments seems to have grown both in the social and biomedical sciences. Very prominent voices in both fields have recently defended the power of randomization. For instance, in 2016, the US Congress passed the 21st Century Cures Act, a bill reforming biomedical research with the goal of bringing new cures more quickly to patients. The Act invited the US Food and Drug Administration to use new trial designs to accelerate the testing process. Among these, pragmatic trials are designed incorporating elements of real-world clinical practice (Ford & Norrie [2016]), like administering treatments in primary healthcare centres, without many of the standard controls that guarantee like with like comparison in conventional RCTs such as blinding. For Robert Califf, then commissioner of the US Food and Drug Administration, randomization provided all the bias control required to guarantee the reliability of pragmatic trials (NAS [2017]). As for economics, the 2019 Nobel Prize was awarded to A. Banerjee, E. Duflo and M. Kremer 'for their experimental approach to alleviating global poverty", and their experimental approach is crucially based on randomized field trials. Our Nobelists claim, for instance: 'In terms of establishing causal claims, it is generally accepted within the discipline that randomized controlled trials are particularly credible from the point of view of internal validity.' (Banerjee *et al*., [2017b], p. 2).

Of course, we are not taking these claims at face value. As we are going to see, the methodological debate on RCTs is still lively both among economists and biostatisticians. Using Worrall's arguments as a thread, we want to explore again when and how randomization can be justified in scientific practice. Our claim is that, in standard statistical practice, balance and control are just instrumental concepts for justifying some forms of statistical inference on experimental data. There is no over-arching concept of balance and control, but, as we are going to show next, there are at least three different traditions, not necessarily consistent between themselves. Worrall

argued that randomization was not necessary to achieve what we will call Millean balance. We will show, on the one hand, that Millean balance is not necessary for solid statistical inferences on experimental data. And that randomization is still necessary for achieving other forms of balance which are still desirable in some widespread statistical approaches to experimental design.[1]

Let us begin with a brief reminder of Worrall's arguments.[2] According to Worrall ([2002]) biomedical researchers have been persuaded by the claim that randomization 'controls for all variables, known and unknown.' RCTs are comparative experiments in which, at least, two groups of participants receive different treatments. In order to make sure that any difference between the observed effects in the two groups originates in the treatments, the groups should be as similar as possible with respect to prognostic factors (balanced). Otherwise, a confounding factor differentially distributed in the two groups may cause the difference between outcomes.[3] It is a widespread view among experimenters that randomization would control for all such confounders, known or unknown. According to some prominent supporters of randomization – cited by (Worrall [2002], pp. 222-223), the strongest argument for this claim would be that a randomized allocation of treatments makes improbable that the distribution of confounders between the two groups 'is very skewed compared to the distribution in the population as a whole', at least if the experiment was repeated for long enough. Worrall objects though that experiments are often run just once, and randomization may just generate a skewed distribution of confounders between the two groups. If the confounders are unknown, there is no way for the experimenters to realize they have been unlucky. Hence randomization does not control for lack of balance.

[1] Just to avoid any misunderstanding, we will not argue that randomization is necessary for any form of causal inference on experimental data.

[2] Worrall ([2002]) considers four different arguments for randomization: significance testing, control of confounders, control for selection-bias, inferiority of observational studies. In addition, Worrall ([2007b]) discusses arguments for randomization in the works of Papineau, Cartwright and Pearl. In our view, Worrall's strongest arguments hinge on the analysis of confounders, and we will take it as the thread for our discussion, introducing his other arguments at different points.

[3] In the literature, there are different ways to name the causal factors other than the intervention influencing the outcome of the experiment. In this paper we will use mainly three: 'antecedent factors' (following J. Stuart Mill), 'confounders' (a common term in the philosophical discussion on causality), and 'covariates' (a standard term among statisticians and trialists). Unless we explicitly signal a nuance, they will be interchangeable.

As we are going to argue next, Worrall's argument seems to presuppose a Millean conception of experimental balance: for causal inference in a comparative experiment to be sound, all the antecedent causal factors (covariates) have a similar value in both groups, so that the intervention is the sole explanans of any difference in the outcome.[4] Following the biostatistician Stephen Senn, we are going to argue that Ronald Fisher's original argument for randomization parted ways with Mill, focusing instead on the analysis of variance. Randomization, for Fisher, did not control for unknown factors guaranteeing a balanced distribution. Instead, it allowed the statistician to measure how big the effect of the intervention was as compared to the effects of all the uncontrolled factors in a given test. The crucial difference is that, for Fisher, the analysis of variance allowed solid causal conclusions even if there was no Millean balance between covariates.

Fisher's approach is, of course, based on a frequentist view of probability and, as Worrall, ([2002], p. 321) contends, no Bayesian will be "persuaded of the need for randomisation, even if it had been convincingly shown that the justification for a significance test presupposes randomization". Drawing again on a recent debate among economists on randomization in field trials, we will argue that, on the one hand, Mill's conception of experimental balance may be inefficient for a Bayesian. Instead of keeping every factor balanced *à la* Mill, a Bayesian experimenter may choose a treatment allocation according to her prior knowledge about potential causes and confounders in order to obtain a better estimate of the treatment effect. The problem is then how to persuade someone who did not share that prior about the convenience of the treatment allocation. In both cases, randomization plays a significant epistemic role, again not as a warrant of Millean balance, but rather in justifying that the treatment allocation process is efficient enough to deliver a good estimate.

The upshot of our analysis is that Worrall is right in showing that randomization does not provide a good warrant of experimental balance in Mill's sense. But for both

[4] Worrall seems to assume that in comparative experiments, randomized or not, it is necessary to control for unbalanced allocations of known factors (Worrall [2002], p. 329; [2007b], p. 486) – see also (Solomon et al. [2017], pp. 202-203). He does not explain how to measure the actual balance achieved, which, in our view, is the crux of the matter. However, our argument is independent of Worrall's position. Since he only targets Millean balance in his arguments, we will show that randomization can be instrumentally defended if the experimenter adopts a different approach, like Fisherian or efficiency balance, about which Worrall remained silent.

frequentist and Bayesian statisticians such understanding of balance is not necessary for causal inference, while randomization is not so easy to dispense with.

In the following three sections, we will introduce three different notions of balance. In section 2, we will present Millean and Fisherian balance: the former is about equal covariate value with *ex ante* control, the latter is about measuring *ex post* the influence of uncontrolled covariates through the analysis of variance. Then, in sections 3 and 4 we will discuss two versions of efficiency balance: the allocation procedure will be assessed according to its contribution to the optimal estimation of the treatment effect. In section 5, we will discuss why scientific progress must presuppose disagreement among experimenters, and this provides a good enough reason for experimenters of all statistical persuasions to keep randomizing.

2 Balance: Mill versus Fisher

John Stuart Mill paradigmatically articulated the concept of experimental balance in his analysis of the Method of Difference. According to Mill, this was the method of 'artificial experiments': the experimenter compares two sets of 'ascertained circumstances', 'resembling one another in every other respect, but differing in the presence or absence of the phenomenon we wish to study' (Mill [1974], p. 386). Those 'ascertained circumstances' are the antecedents of the phenomenon under study: the comparison should show which of them is its proper cause, the factor that suffices to produce the effect when present, and whose absence makes it disappear. The rest of antecedents would just be mere conditions.

For Mill, an experimental comparison is balanced if the sets of antecedent circumstances, other than the putative cause, are exactly alike. He was, of course, aware of the difficulties. Assessing the effects of a medical treatment in an experiment is difficult because there are so many antecedent causes contributing to the effect that the experimenter will rarely succeed in separating them from the actual intervention. Unknown confounders can only be ruled out if the experiment is carried out under so many different circumstances that it becomes unlikely that a given set of unknown confounders is at work in all the comparisons. Even known confounders are difficult to control for if they interact with the intervention under study to produce the effect. For

Mill, comparative experiments in medicine are only conclusive if the intervention is 'more potent than any counteracting causes' (Mill [1974], p. 451), so that they succeed in restoring health in a large number of cases. But such powerful interventions are rare and for regular treatments medical experiments usually fail to establish causality for lack of balance (Mill [1974], p. 451).

Although Mill's approach has been extensively criticized, his notion of experimental balance is still widespread in philosophy of science. As Hofmann & Baumgartner ([2011]) put it, 'the standard opinion in the literature, from Mill to Woodward, has been that under homogeneous experimental conditions, i.e. when possible confounders of an investigated deterministic structure are controlled, a single positive difference-test result is sufficient for a causal inference.' Along different lines, Hofmann & Baumgartner ([2011]) and Scholl ([2015]) have argued that balancing potential confounders does not suffice to support causal inference through Mill's method of difference –see also (Fuller [2019]).

Following Stephen Senn ([2013]), we are going to argue next that Ronald Fisher, one of the founders of modern statistical inference, took a different approach to experimental balance in causal inference. For Fisher, Millean balance is not even necessary for causal inference. Like Mill, Fisher argued that 'it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable.' (Fisher [1971], p. 18) This endless list of confounders was, for Fisher, a source of error in the estimate of the effects of a treatment. Unlike Mill, Fisher used comparative experiments to quantify the contribution of this error to the observed effect (Hall [2007]).

Fisher saw how the uncontrolled causes would introduce variation in each arm of the experiment: a bigger or smaller range of values of the outcome variable measuring the effects of a treatment. Fisher's insight was to compare the amount of variation within each treatment group and the amount of variation between groups (the difference between the average treatment effects in each of them). If the difference was statistically significant, then the experimenter could conclude that the intervention is "more potent than any counteracting causes". Unlike Mill, Fisher's analysis of variance did not try to control for the *ex ante* value of each possible confounder, but for their

aggregate effect *ex post*. Causal inference in well-designed experiments did not depend on Millean balance, but on the proper statistical interpretation of the outcome.

In Fisher's approach, randomization does not contribute to balance, but rather to assess the statistical significance of the observed difference between treatments. Comparative experiments are conducted under the (null) hypothesis that there is no difference between treatments, and that if any difference is observed it will be due to uncontrolled factors. Randomizing the allocation of treatments allowed Fisher to quantify the statistical significance of each observed difference between treatments. Under the null hypothesis, between- and within- group variance would, on average, be identical. Since the probability of each randomized allocation was known, Fisher could calculate how likely it was to observe a given range of differences between treatments, including the observed value. And then decide whether this observed value meant that something unusual has happened, or rather that the null hypothesis is false (there is an actual difference between the treatments).[5]

For Mill balance was, almost always, a prerequisite for causal inference in experiments. If the experimenter had not enough control of known and unknown confounders to achieve it, causal inference was only possible when the intervention was powerful enough to counteract them. For Fisher, randomization gave the experimenter control on all the unknown confounders. For controlling the known confounders, Fisher advised gathering them into blocks and then randomizing the treatment within blocks.[6] But blocking only increased the precision of the estimate: lack of balance between known confounders did not make the estimation less valid, only less precise. Let us illustrate it with Senn's ([2020]) own example.

Let us compare two trials with different degrees of Millean balance. On the one hand, there is Trial A, a cross-over trial in which each patient sequentially receives just

[5] Significance tests, on their own, provide just evidence of correlation, not causation. The experimenter's causal knowledge informs the design of the test: defining the intervention, blocking known confounders etc. Significance tests just provide a device to interpret the outcome. But we are not comparing Mill's versus Fisher's inference methods, but rather the role balance plays in their experimental designs.

[6] Now, what if randomization generates a Millean imbalanced allocation, where, to the naked eye, one of the antecedent factors is unevenly distributed? According to the witness testimony of William Cochran (Rubin [2008]), if the experiment had not started, Fisher would rerandomize. Fisher never justified rerandomization in print, and, as we will see below, there are different options to do so –see (Savage [1976], p. 464) for further historical details on Fisher. The point is, against (Urbach, 1985), that rerandomizing for Millean balance is not necessary to conduct the analysis of variance, but rather to achieve a more precise estimate of the treatment effect –see section 3 below.

one treatment in order to compare their effects. Even if the order of treatment administration is randomized, this trial achieves a high degree of Millean balance, since the relevant antecedent factors such as genes are the same for each patient. On the other hand, the same two treatments are tested in a randomized parallel design in which each patient only receives one of them. This is Trial B, and here there is no control for Millean balance: the relevant covariates may have different average values.[7]

In Trial A there are 71 patients, yielding 142 observations. In Trial B, there are 37 patients, yielding 74 observations. Trial A is much more precise than Trial B: the 95% confidence interval of the former for the variable estimating the treatment outcome is [0.1, 0.23]. whereas the latter is [0.02, 0.74]. The higher precision of Trial A is due to both sample size and, crucially, to Millean balance (each patient being her own control). But Trial B is nonetheless statistically valid, only less precise: there are just different degrees of uncertainty to the conclusion, broader or narrower confidence intervals. In both cases, randomization allows Fisherians to compare the effects of both treatments and draw a solid conclusion, provided that they use the appropriate test for the degree of balance in each trial – a matched pair t-test for Trial A and a two sample t-test for Trial B.

Therefore, in a Fisherian approach, randomization does not contribute to attaining Millean balance in RCTs: blocking does. As Stephen Senn ([2013]) forcefully argued, against Worrall, statistical significance can be assessed in any single RCT without assuming any long run view about the balancing effects of randomization. Indeed, randomization guarantees that, averaged over an infinite number of replications of the test, the error in estimating the treatment effect will be zero. But in any single run of the experiment, randomization allows the statistician to calculate the probability of outcomes as big or more than the actually observed result, under the assumption that there is no difference between treatments -any variation will have arisen from uncontrolled factors derived from treatment allocations (Basu [1980]). For Fisher, the experimenter should decide whether a statistically significant event implies that the null

[7] In economics, designs of the type of trial B are more abundant than Trial A, because the interventions often require many months to observe medium-run effects – think of means-tested transfer programs such as the Earned Income Tax Credit (Nichols & Rothstein [2015]), or of various configurations of a Basic Income experiment (Hoynes & Rothstein [2019]).

hypothesis is false (there is an actual difference between treatments) or, rather, that there are confounders at work and the experiment should be repeated. If the statistically significant outcome vanishes in further replications, this second option will be justified For Fisher, the experimenter only had a 'real phenomenon' under control when she could repeat the experiment time and again and rarely fail to obtain a statistically significant outcome (Spanos & Mayo [2015]).

This is, of course, a fallible decision, and Fisherian *p*-values have been extensively criticized as decision criteria (Sprenger [2016]). But we are not trying to vindicate *p*-value here, we just want to clarify how Fisher's approach to experimental balance in causal inference is different from Mill's. In our view, Worrall is criticizing standard RCTs as if their method presupposed something like Millean balance, when it does not. Still, Worrall pushes forward and contends that balance can be attained from a Bayesian perspective in which randomization is not necessary. This is the claim that we are going to target next:

> Once you have made sure that there is no positive reason to think the two groups are unbalanced (and this automatically means checking for imbalance in factors you know about), then whether or not the division was produced by following some table of random numbers or tossing a fair coin, or just by happenstance, can be of no epistemic account. This is what the Bayesian is saying, and it seems to me entirely convincing. (Worrall [2007b], p. 466)

3 Balance in a Bayesian Perspective: Efficiency Balance

From a Bayesian perspective, randomization is not a pre-requisite for interpreting the outcome of any comparative experiment (like clinical trials). A Bayesian design will set some prior probabilities about the outcome, run the test and update those priors in the light of the actual outcome. For this updating process, a Bayesian does not need to know the probability (*p*-value) of observing the actual (or a more extreme) outcome, so randomization, in this regard, becomes dispensable.

Bayesian experimenters may have other reasons to randomize, and some of them have been discussed decades ago: for instance, Kadane & Seidenfeld ([1990]) or

Berry & Kadane ([1997]) show that an experimenter should randomize in order to make the test outcome credible for third parties: e.g., if the person running the experiment is different from the person conducting the statistical analysis of the outcome and the latter doesn't trust the former. Randomization would guarantee the neutrality of the allocation regarding the interests of the experimenter- we will return to this point later. We are now going to analyse the different notions of balance at stake in a recent methodological debate among economists, and the role randomization plays in articulating these notions. As we interpret it, the upshot of this debate is that, from a Bayesian perspective, randomization might be necessary to achieve experimental balance. Except that now balance will be interpreted in terms of efficiency in the estimation of the treatment outcome. Let's call this notion *efficiency balance* and let us show how it is different from Fisherian and Millean balance.

An experimental design is efficient depending on the sample size required to estimate the effect of a treatment. This estimation is usually assessed in terms of bias and precision. Randomization provides a warrant of unbiasedness (average difference between the estimator and the true value): in the long run, the error term in the estimate of the treatment effect (the sum of the net average balance of other causes across the two groups in a trial) will be zero. However, randomization may have an impact on precision (how close to the truth is the estimator on average): in any single run of the experiment, a randomized allocation of treatments may generate an imbalanced distribution of covariates, shifting away the estimator from the true treatment effect. As Deaton & Cartwright ([2018], p. 5) put it, Fisherian balance is only acceptable if the experimenter is willing to sacrifice truth for the sake of unbiasedness.[8]

An efficient experimental balance is achieved through an allocation that controls for covariates in a way that minimize bias and maximizes the precision of the point estimate. The efficiency of Fisherian balance can be improved through blocking. In clinical trials, restricted forms of randomization (such as stratification or minimization) are well-known strategies to control *ex ante* for baseline covariate imbalance (Senn

[8] However, Deaton and Cartwright ([2018], p. 6) acknowledge that the virtue of randomization in a Fisherian approach, is 'getting the standard error and associated significance statements right'. As a reviewer observes, for Fisher, the quality of a statistical analysis does not lie in the precision of the point estimate alone ('truth'), but on how the probability of error in such estimate is quantified via significance statements etc. For the latter, randomization is still necessary.

[2007]). But they are often difficult to implement, leaving unrestricted randomization as the default procedure (Ciolino *et al*. [2019]).

The situation is different in field trials testing in economics. Again, these are comparative experiments in which, at least, two policy interventions are tested to see which one is more effective in bringing about the desired policy outcome. Whereas in most clinical trials patients are enrolled in the test sequentially (e.g., with the onset of their symptoms), in economic experiments it is possible to randomize and study potential imbalances before the actual start of the experiment. The experimenters should agree on a list of relevant covariates that is necessary to control for. After randomizing, they should then check whether these covariates have, on average, a similar enough value.[9] In field trials in economics, there are different *ex post rules of thumb* to assess the significance of certain imbalances in an experiment (Bruhn & McKenzie [2009]). For instance, trialists take a random draw from the randomly allocated treatments, and then check the difference in means for some key covariates. If the difference looks too large, then the standard fix is to re-randomize the allocation.

As Imbens and Rubin ([2015], p. 81) put it, in most situations 'researchers are not solely interested in obtaining p-values for sharp null hypotheses. Simply being confident that there is some effect of the treatment for some units is not sufficient to inform policy decisions.' Therefore, economists rely on regression-based approaches on balanced samples trying to capture the widespread effect of the treatment on the target population. Balance here is understood in a quasi-Millean fashion: the relevant covariates in the two groups should not be too different, on average Mill did not think of equality between factors in probabilistic terms. But this quasi-Millean balance is not a prerequisite for causal inference via regression analysis, while randomization is. Randomization guarantees that the treatment variable is statistically independent of unknown confounders that could affect the outcome directly or through the treatment variable. Thus, Fisherian balance is all regression analysis needs to reach solid causal conclusions. Quasi-Millean balance is just a desirable feature to have for two reasons. First, as Imbens and Rubin ([2015], p. 114) claim, 'if the covariates are predictive of the

[9] As a reviewer observes, 'this procedure would not be regarded as valid by Fisher and his followers. Ignoring covariates if they look sufficiently balanced does not lead to valid inference. Calculation of standard errors will not be correct.' See (Senn [2008]) for some examples and discussion.

potential outcomes, their inclusion in the regression model can result in causal inferences that are more precise than differences in observed means [. . .] although in practice the gains are often modest'.[10] Second, it provides a justification for extrapolating the test outcome to the target population outside the experiment.

However, a Bayesian experimenter can entirely dispense with randomization, and therefore with Fisherian balance, focusing entirely on efficiency. Kasy [2016]) provides a paradigm for this approach. For Kasy, balance as equal distribution of covariates is too demanding an ideal: such distribution is rarely identical between the treatment and control groups. The experimenter needs to trade off balance across the various dimensions of the joint distribution of covariates and the question is how to justify this trade-off in a systematic manner.

From a purely Bayesian perspective, this becomes a decision problem for the experimenter. She should design the trial, choosing a treatment allocation procedure and the estimator, in the light of the covariate distribution in the sample. For Kasy, the experimenter should minimize the conditional expected loss function of an estimator, representing the risk of a difference between estimated and true values.[11] A treatment allocation will be then balanced to the extent that minimizes that loss function. This is efficiency balance.

For Kasy, adopting a prior over the potential data-generation processes and a tractable loss function (the mean squared error, MSE), it is possible to construct an optimal allocation procedure solving the experimenter's decision-theoretic problem.[12] Briefly, Kasy suggests to randomize the allocation a pre-established number of times, picking up the assignment that minimizes most the loss function. Randomization plays here no inferential role: it is just an impartial device for choosing the assignment.

[10] Conversely, see (Senn *et al.* [2010]) for a discussion of efficiency in medical trials, showing that randomization does not significantly diminishes efficiency.

[11] Since the decision about the estimator is made before observing the actual outcomes of the experiment, the choice should hinge on its expected loss.

[12] Bias and variance are the two components of the MSE: for bias = 0, the MSE is the variance. But Kasy correctly observes that experimental design proceeds without knowledge of the underlying data generating process. Hence, Kasy suggests to use the expected MSE, averaging the MSE over possible data generating processes. He uses a nonparametric Bayesian prior over those processes to construct his allocation procedure, which is to randomize k times the allocation of treatments, picking the one with the best MSE –assuming that prior.

For Kasy, randomization is otherwise dispensable. A randomized allocation is just a random pick from the set of all possible treatment allocations. Each one of these allocations will exhibit a particular distribution of covariates, generating a particular value for the experiment's loss function (the MSE). The risk of a randomized allocation is just the weighted average of the mean squared error across all the treatment assignments it averages over. A deterministic allocation rule that picks up a particular treatment assignment among all those that make the mean squared error minimum will dominate any randomization scheme, because the estimator will have a lower risk.

Since finding the allocation providing the minimal value for the loss function generally is an intractable task, Kasy suggests as a shortcut his randomized procedure. Under certain assumptions, equality between covariate means may minimize a given loss function, but this is just a particular implementation of efficiency balance. An experiment with a Millean imbalanced allocation may, nonetheless, yield a good estimate and have efficiency balance.

Defining balance in terms of efficiency provides a more systematic justification of a particular covariate distribution that Fisher's blocks or Mill's equality between factors. However, as we are going to see next, this justification comes at a price: it is unpersuasive if experimenters disagree on their priors.

4 Agreeing on Covariates

We have discussed so far three notions of balance. Millean balance is a pre-statistical notion targeting the single run experiment in which all confounding factors are kept at the same value. Fisherian and efficiency balance are statistical concepts. In the former, control of imbalances is achieved comparatively and *ex post* through the analysis of variance, with optional *ex ante* control via blocking. In the latter, control of imbalances is achieved *ex ante*, with the experimenter drawing on her prior knowledge of the relevant covariates to minimize a loss function.

The choice between these three competing notions of balance depends, crucially, on how the experimenter understands causal and statistical inference. When John Stuart Mill articulated his concept of balance, he was mostly unpersuaded by statistical approaches; in all likelihood, frequentists and Bayesians will fail to agree on the concept of balance due to their more fundamental differences. But leaving aside

those principled disagreements, there is one fundamental question about which experimenters must agree, whatever their concept of balance: which factors (covariates, known confounders) should they control for balance?

Sometimes theories dictate which are the relevant factors to control for in an experiment. It is indeed an ongoing controversy in economics whether field trials can identify causal structures independently of any theory.[13] But even when such theories exist and identify the causally relevant variables, experimenters often need to rely on their informal knowledge of potential confounders present in the field but not covered by the theory that should be, nonetheless, controlled for. This is what an economist *as a plumber* should do, as Duflo ([2017]) puts it. The division of labour is as follows. First, the economist as scientist defines the broad program design structure according to the relevant scientific evidence and theories. Then, the economist should wear a plumber's hat and ask which specific details of the context where the program is implemented could affect the program effectiveness in the field context of interest. If necessary, she should reshape the design of the interventions accordingly. In any case, such contextual details will matter in ruling out potential confounders.

For example, Duflo ([2017]) discusses programs that are funded centrally but implemented at the local level, such as the Raskin Indonesian rice distribution scheme. Building on a literature that emphasizes the leakages occur in foreign aid and governmentally supported programs, Banerjee and his co-authors ([2015]) want to test to what extent transparency diminishes those leakages. They design their experiments around interventions in which citizens are exposed to different degrees of information about the rice distribution program. The authors then used local knowledge, acquired in the field, to tailor the details of the program to the Indonesian context. For example, they partner with the central government rather than with the local administration for the implementation, after having observed the degree of discretion that local officers had in deciding the amounts of rice to distribute. The experimenters even quantified by how much the details mattered in program effectiveness.

[13] There is an ongoing controversy in economics as to whether field trials should be theory-free or they should instead draw some causal assumptions from structural models, but we will leave it aside here – see (Banerjee & Duflo [2010]) or (Boumans [2016]) for further discussion.

The epistemic question is how the experimenters should agree on which covariates to balance, when there is no consensus on their relevance. In a Bayesian approach like Kasy's, this question amounts to whether a particular prior about these covariates will persuade other experimenters who do not share it. Picking up this thread, Banerjee and co-authors defend the use of randomization as a balancing device for its ability to persuade audiences, whatever their view of probability. This is what they call adversarial experimentation: the experimenters in a community have different priors about the relevant covariates to control for.

Adopting again a decision-theoretic perspective, Banerjee and co-authors ([2017]) analyse field experiments testing policy interventions. The experimenter should choose an experimental design and a decision rule about the policy to implement, in the light of the outcome. The experimenter should maximize here a payoff function with two components. On the one hand, there is the expected subjective utility of a decision rule, given the experimenter's own prior. On the other hand, there is the minimal value that the same expected utility function for all other priors in the community. In order to maximize the payoff function, the experimenter should trade off her own persuasion with the minimal amount of persuasion her choice would generate in other members of the community. Banerjee and co-authors show that randomization is dispensable if the experimenter cares most about her own persuasion, but not if she cares most about convincing the community.

Banerjee and co-authors' proof hinges crucially on the assumption that, in adversarial experiments, there is always one prior such that, for any non-random treatment allocation, the experimenter holding that prior won't be completely persuaded about the correct policy choice (Banerjee *et al*. [2020]). Whereas Kasy ([2016]) uses a single standardized prior to construct a deterministic allocation rule minimizing the loss function, expecting that the community of experimenters will agree on the convenience of this particular prior. Banerjee and co-authors assume instead an actual diversity of priors in the communities of experimenters. If the allocation rule is predictable, an experimenter unwilling to accept the outcome can always construct her prior in a way that she will remain unpersuaded. Randomizing the allocation prevents such strategic choice of priors.

Moreover, Banerjee and co-authors challenge Kasy's approach to efficiency in terms of persuasion. Different treatment allocation rules may achieve different degrees of efficiency measured as deviations from the first best decision, which is the experimenter's own payoff function, disregarding the audience). Banerjee and co-authors prove that there is an upper bound to the efficiency loss of the optimal experiment as compared to the first best: experimenters' do not lose much precision if they opt for randomizing. At the same time, Banerjee and co-authors. show that no deterministic rule is optimal (although they may be a first best for an experimenter).

In other words, efficiency balance *à la* Kasy is only persuasive if there is a consensus on the prior for relevant covariates. Where does this leave us?

5 Randomization and the Progress of Experiment

There are, at least, two different approaches to balance in which there are reasons to randomize. Those who side with Fisher, seek balance blocking for known confounders and measuring the interferences of unknown confounders through the analysis of variance, for which randomization is necessary. Bayesians can codify in their priors what they know about potential covariates and justify their treatment allocations in terms of the efficiency of their estimates. But if there are different priors about those covariates, a randomized allocation will yield a consensual and efficient enough estimate.

In both cases, randomization and balance are instrumental concepts. In the Fisherian approach, the balance achieved through randomization is a tool for obtaining a reliable estimate of the imprecision arising from a potentially endless list of covariates. In the efficiency approach, the balance achieved through randomization (either for the single experimenter *à la* Kasy, or for the community) justifies that the covariates included in the model are the relevant ones for an efficient estimation of the treatment effect.

As a reviewer observed, the upshot of this analysis is that, from the standpoint of the statistical design of experiments, balance is, in fact, a red herring. The experimenter should focus on the relevant statistical indexes tracking the quality of the data analysis (standard errors, efficiency). As we have seen, depending on the experimenter's view of the data analysis, she will adopt derivatively one or another

concept of balance. We remain agnostic here about the ultimate goals the data analyst should pursue. We are just showing how randomization contributes to achieving them.

Worrall's objections against randomization as a warrant of balance seem to be based on a purely Millean approach, the only one targeted in his arguments. This Millean approach seems just rough: equalizing covariates between groups in trials is often hard and, on its own, does not contribute crucially to the statistical estimation of the treatment effect. Worrall will surely acknowledge that there can be solid causal inference, in both a frequentist and a Bayesian approach, without anything like Millean balance. The price the experimenter pays for not having it is, at most, lack of precision in the estimate of the treatment outcome. We have argued that lack of precision is a reasonable price to pay in adversarial experimentation, if randomization brings about consensus among experimenters.

Of course, trading off precision and consensus depends on our take on scientific progress. If adversarial experimentation were more the exception than the rule in science, randomization might be, in the end, dispensable. Deaton and Cartwright had pushed this point in their recent paper: 'The systematic refusal to use prior knowledge and the associated preference for RCTs are recipes for preventing cumulative scientific progress' (Deaton & Cartwright [2018], p. 7). Although they acknowledge that randomization is the lesser evil when experimenters disagree about priors, it should not be the default mode for the advancement of experiments.

In our view, this is a misleading claim. On the one hand, prior knowledge may be incorporated into both frequentist trials (via blocking) and into Banerjee's approach via Bayesian priors. On the other hand, we should not assume that cumulative scientific progress presupposes agreement among experimenters on the relevant covariates. This may be the case in normal science, but eventually they will bump into Kuhnian anomalies prompting them to disagree on their controls. Historians and sociologists of experimentation have shown how the elimination of background confounders is a key part of the process leading to a renewed agreement on experimental phenomena, and bias control is a central ingredient of this discussion –e.g., (Collins [1981]; Galison [1987]; Teira [2013]). When experimenters disagree their conflicts of interest might not be motivated by the sort of financial concerns that pervade regulatory trials in medicine, but they are no less real. As the replication crisis in psychology has illustrated, the simple

desire for professional promotion or public recognition is enough for researchers to fiddle with experimental designs.

Still, it may be argued that even if adversarial experimentation is frequent in science, randomization is not the best control for such biases. Worrall ([2002]) granted that randomization was an effective control for one major source of experimental fiddling: selection bias, the experimenter influencing the trial outcome through an intentional allocation of treatments. But he downplayed its epistemic value as follows:

> Notice however that randomization as a way of controlling for selection bias is very much a means to an end, rather than an end in itself. It is blinding (of the clinician) that does the real methodological work—randomization is simply one method of achieving this. (Worrall [2002], p.325)

It is open to discussion whether blinding can really work without randomization (Senn, [2013]), but the problem in deciding what covariates to control for and how to do it is that it cannot be done in the blind. As we already saw in sections 3-4, selection bias can affect the trial outcome through covariate selection just as much as through treatment assignment. When there is no solid scientific consensus about which covariates should be balanced and how should this balance be achieved, the experimenters are still better off randomizing, for lack of a better alternative.

Summing up, as Fisher himself once put it, 'whatever degree of care and experimental skill is expended in equalising the conditions, other than the one under test, which are liable to affect the result, this equalisation must always be to a greater or less extent incomplete' (Fisher [1971], p. 19). We have tried to show in this paper that there are at least three competing notions of covariate balance in scientific experimentation. Worrall's arguments targeted a Millean conception of balance. For, practising scientists, like Fisher or Kasy, complete balance is unattainable, and it is necessary for experimenters to agree on which forms of imbalance are acceptable in their tests. As we have tried to argue here, achieving this consensus on balance still provides a good enough reason to randomize.

Acknowledgements

Thanks to our anonymous reviewers, Donald Gillies, Adam La Caze, Julian Reiss and Stephen Senn for their extensive comments on our drafts. This paper also benefited from feedback by Anna Mergoni, Anna Alexandrova, Charlotte Sophia Bez, Francesca Chiaromonte, Rebecca Livernois, and Alessio Moneta and the participants in the 2019 Lake Como Summer School on Economic Behaviours. All remaining errors are our own. Funding for this research was provided by the Spanish Ministry of Science (RTI2018-097709-B-I00 to David Teira).

Marco Martinez
Scuola Superiore Sant'Anna
Pisa, Italy
Marco.martinez@santannapisa.it

David Teira Serrano
UNED
Madrid, Spain
dteira@fsof.uned.es

References

Abdul Latif Jameel Poverty Action Lab (J-PAL) [2020]: 'The Abdul Latif Jameel Poverty Action Lab', available at https://www.povertyactionlab.org

Angrist, J. D. and Pischke, J.-S. [2010]: 'The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics', Journal of Economic Perspectives, 24, pp. 3–30.

Banerjee, A., Hanna, R., Kyle, J. C., Olken, B. A. and Sumarto, S. [2015]: 'The Power of Transparency: Information, Identification Cards and Food Subsidy Programs in Indonesia', National Bureau of Economic Research Working Paper Series, available at http://www.nber.org/papers/w20923

Banerjee, A. V. and Duflo, E. [2010]: 'Giving Credit Where It Is Due', *Journal of Economic Perspectives*, **24**, pp. 61–80.

Banerjee, A. V., Chassang, S. and Snowberg, E. [2017]: 'Decision Theoretic Approaches to Experiment Design and External Validity', in E. Duflo and A. Banerjee (eds.), *Handbook of Economic Field Experiments*, Vol. 1, Amsterdam: Elsevier, pp. 141–74.

Banerjee, A. V., Chassang, S., Montero, S. and Snowberg, E. [2020]: 'A Theory of Experimenters: Robustness, Randomization, and Balance', *American Economic Review*, **110**, pp. 1206–30.

Basu, D. [1980]: 'Randomization Analysis of Experimental Data: The Fisher Randomization Test', *Journal of the American Statistical Association*, **75**, pp. 585–81.

Berry, S. M. and Kadane, J. B. [1997]: 'Optimal Bayesian Randomization', *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**, pp. 813–9.

Boumans, M. [2016]: 'Methodological Ignorance: A Comment on Field Experiments and Methodological Intolerance', *Journal of Economic Methodology*, **23**, pp. 139–46.

Bruhn, M. and McKenzie, D. [2009]: 'In Pursuit of Balance: Randomization in Practice in Development Field Experiments', *American Economic Journal: Applied Economics*, **1**, pp. 200–32.

Ciolino, J. D., Palac, H. L., Yang, A., Vaca, M. and Belli, H. M. [2019]: 'Ideal vs. Real: A Systematic Review on Handling Covariates in Randomized Controlled Trials', *BMC Medical Research Methodology*, **19**, p. 136.

Collins, H. M. [1981]: 'Son of Seven Sexes: The Social Destruction of a Physical Phenomenon', *Social Studies of Science*, **11**, pp. 33–62.

Deaton, A. and Cartwright, N. [2018]: 'Understanding and Misunderstanding Randomized Controlled Trials', *Social Science & Medicine*, **210**, pp. 2–21.

Duflo, E. [2017]: 'Richard T. Ely Lecture: The Economist as Plumber', *American Economic Review*, **107**, pp. 1–26.

Fisher, R. A. [1971]: 'The Design of Experiments', 8th ed., New York: Hafner Publishing Company.

Ford, I. and Norrie, J. [2016]: 'Pragmatic Trials', *New England Journal of Medicine*, **375**, pp. 454–63.

Fuller, J. [2019]: 'The Confounding Question of Confounding Causes in Randomized Trials', *The British Journal for the Philosophy of Science*, **70**, pp. 901–26.

Galison, P. [1987]: 'How Experiments End', Chicago: University of Chicago Press.

Hall, N. S. [2007]: 'RA Fisher and His Advocacy of Randomization', *Journal of the History of Biology*, **40**, pp. 295–325.

Hofmann, U. and Baumgartner, M. [5]: 'Determinism and the Method of Difference', *THEORIA. An International Journal for Theory, History and Foundations of Science*, **26**, pp. 155–76.

Hoynes, H. and Rothstein, J. [2019]: 'Universal Basic Income in the United States and Advanced Countries', *Annual Review of Economics*, **11**, pp. 929–58.

Imbens, G. W. and Rubin, D. B. [2015]: *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Kadane, J. B. and Seidenfeld, T. [1990]: 'Randomization in a Bayesian Perspective', *Journal of Statistical Planning and Inference*, **35**, pp. 329–45.

Kasy, M. [2016]: 'Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead', *Political Analysis*, **24**, pp. 324–38.

Mill, J. S. [1974]: *A System of Logic, Ratiocinative and Inductive, Books I-III*, 3d ed., Toronto: University of Toronto Press.

NAS (The National Academy of Science) [2017]: *Real-World Evidence Generation and Evaluation of Therapeutics: Proceedings of a Workshop*, Washington (DC): National Academies Press (US).

Nichols, A. and Rothstein, J. [2015]: 'The Earned Income Tax Credit', in R. A. Moffit (ed.), *Economics of Means-Tested Transfer Programs in the United States,* Chicago: University of Chicago Press, pp. 137–218.

Rubin, D. B. [2008]: 'Comment: The Design and Analysis of Gold Standard Randomized Experiments', *Journal of the American Statistical Association*, **103**, pp. 1350–3.

Savage, L. J. [1976]: 'On Rereading R. A. Fisher', *The Annals of Statistics*, pp. 441–500.

Scholl, R. [2015]: 'Inference to the Best Explanation in the Catch-22: How Much Autonomy for Mill's Method of Difference?', *European Journal for Philosophy of Science*, **5**, pp. 89–110.

Senn, S. [2007]: 'Statistical Issues in Drug Development', 2nd ed., Chichester, England ; Hoboken, NJ: John Wiley & Sons.

Senn, S. [2008]: 'Lessons from TGN1412 and TARGET: Implications for Observational Studies and Meta-Analysis', *Pharm Stat*, **7**, pp. 294–301.

——— [2013]: 'Seven Myths of Randomisation in Clinical Trials', *Statistics in Medicine*, **32**, pp. 1439–50.

——— [2020]: 'Randomisation Is Not about Balance, nor about Homogeneity but about Randomness', available at Error Statistics Philosophy Blog, https://errorstatistics.com/2020/04/20/s-senn-randomisation-is-not-about-balance-nor-about-homogeneity-but-about-randomness-guest-post/.

Senn, S., Anisimov, V. V. and Fedorov, V. V. [2010]: 'Comparisons of Minimization and Atkinson's Algorithm', *Statistics in Medicine*, **29**, pp. 721–30.

Solomon, M., Simon, J. R. and Kincaid, H. (eds.) [2017]: *The Routledge Companion to Philosophy of Medicine*, New York: Routledge, Taylor & Francis Group.

Spanos, A. and Mayo, D. G. [2015]: 'Error Statistical Modeling and Inference: Where Methodology Meets Ontology', *Synthese*, **192**, pp. 3533–55.

Sprenger, J. [2016]: 'Bayesianism vs. Frequentism in Statistical Inference', A. Hájek and C. Hitchcock (eds.), *Oxford Handbook of the Philosophy of Probability*, Oxford: Oxford University Press, pp. 382–405.

Teira, D. [2013]: 'A Contractarian Solution to the Experimenter's Regress', *Philosophy of Science*, **80**, pp. 709-720.

Urbach, P. [1985]: 'Randomization and the Design of Experiments', *Philosophy of Science*, **85**, pp. 256–73.

Worrall, J. [2002]: 'What Evidence in Evidence-Based Medicine?', *Philosophy of Science*, **69**, pp. S316–30.

——— [2007a]: 'Evidence in Medicine and Evidence-Based Medicine', *Philosophy Compass. N*, **2**, pp. 981–1022.

——— [2007b]: 'Why There's No Cause to Randomize', *British Journal for the Philosophy of Science. S*, **58**, pp. 451–88.

——— [2008]: ‘Evidence and Ethics and Medicine’, *Perspectives in Biology and Medicine*, **51**, pp. 418–31.